Universidad de Jaén

Escuela de Doctorado

DOCTORAL THESIS

•

Detecting offensive language by integrating multiple linguistic phenomena

PRESENTED BY:
Flor Miriam Plaza-del-Arco

SUPERVISED BY:

PhD. María Teresa Martín-Valdivia
PhD. L. Alfonso Ureña-López

JAÉN, November 2022

"Per aspera ad astra"

- Lucio Anneo Séneca

Abstract

Digital technologies have transformed the way people communicate, turning the Web into a global means of communication in our daily lives. Since the advent of social media, more and more people are expressing their opinions and sharing their experiences. However, this expression does not always create a healthy environment; rather, it can occasionally encourage users to act in a harmful attitude, which is sometimes aided by the anonymity that these platforms allow. Online users may experience negative psychological impacts from this form of hostile communication, including anxiety, harassment, and, in severe cases, suicidal thoughts. As a result, this situation has motivated governments and online content moderators to search for efficient solutions to prevent Internet hostility by implementing laws and policies. However, the types of strategies adopted are not sufficient, since they involve an intense, time-consuming, and costly procedure that limits scalability and quick solutions. Natural language processing, one of the primary disciplines of Artificial Intelligence, is crucial to combat this situation and offensive language detection and analysis has become a major area in this field. This doctoral thesis focuses on automatic offensive language detection by the generation of linguistic resources and the development of automatic NLP-based methods. Firstly, we tackle the problem of data scarcity, especially in Spanish. We present a lexicon resource and three different corpora along with benchmarks to validate them. In order to promote the research in this area, we organize different shared tasks using the resources generated. Secondly, we propose different linguistic phenomena that could be involved in the expression of offense. Then, we develop a novel methodology that takes advantage of the transfer learning paradigm to integrate these phenomena. Results show an increased performance of our proposed method over state-of-the-art systems. Thirdly, this novel method is applied to different scenarios of offensive language, analyzing which specific linguistic phenomena are beneficial in each of these scenarios. Finally, we summarize our contributions and suggest for future research directions on the offensive language research area.

Contents

List of Figures

List of Tables

Abbreviations

General notation

e.g.	**e**xemplum **g**ratia (*en*: for example)
et al.	**et al**ia (*en*: and others)
i.e.	**i**d **e**st (*en*: that is)

Machine learning

ML	**M**achine **L**earning
MSE	**M**ean **S**quared **E**rror

Deep learning

DL	**D**eep **L**earning
MTL	**M**ulti-**T**ask **L**earning
NN	**N**eural **N**etwork

Natural language processing

AI	**A**rtificial **I**ntelligence
NLP	**N**atural **L**anguage **P**rocessing
LM	**L**anguage **M**odel
MT	**M**achine **T**ranslation
NER	**N**amed **E**ntity **R**ecognition
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requency

To my parents, Joaquín & Marci, and my sister, Ana Belén.

Chapter 1

Introduction

One of the characteristics that distinguish humans from other living beings is the ability to communicate in a systematic and understandable manner, i.e. through language. Language is defined as a sophisticated system of both phonetic and written symbols that allows two or more individuals to communicate ideas, thoughts, sentiments, attitudes, and different situations. Since the emergence of Web 2.0, users were no longer limited to face-to-face communication but rather used online platforms to interact. This interaction has resulted in an increasing amount of textual data being available on the Web and therefore, the NLP, a tract of Artificial Intelligence and Linguistics, arises for the development of computational systems to interpret human language and thus enable human-computer interaction. Giving computers this skill offers a plethora of benefits, including the potential to moderate harmful conduct on social media.

This doctoral thesis focuses on both the creation of linguistic resources and the development of NLP-based techniques to aid in the automatic detection of offensive language on the Web. On the one hand, for the development of these techniques, data labeled are essential to learning the language patterns characteristic of this behavior; however, the available resources are mainly focused on English, leaving aside other languages such as Spanish with very scarce or non-existent resources of this nature. Therefore, a fundamental part of this doctoral thesis is focused on the generation of these resources for Spanish. On the other hand, for the implementation of automatic systems based on NLP, one of the main ideas generated has been the integration of different linguistic phenomena that can be involved in the expression of offensiveness in the computational systems. We believe that this methodology plays an important role in their application to the detection of more specific problems in our society, such as Hate Speech (HS), misogyny, or sexism problems that have been addressed in the frame

of this doctoral thesis. As a result, it should be mentioned that this thesis has both a social and technological dimension to contribute to society's improvement.

1.1 Motivation

Social media have grown into the primary means of communicating between people, allowing users to have conversations, share opinions and create content. The rise in digital social connections has led to the dissemination of harmful communication, which is sometimes aided by the anonymity afforded by these platforms [2]. As a consequence, offensive language and one of its most damaging forms, HS, has the tendency to proliferate swiftly and is difficult to regulate. For instance, according to a Spanish report in 2020 on the evolution of hate crimes in Spain[1], threats, insults, and discrimination are counted as the most repeated criminal acts, with the Internet (45%) and social media (22.8%) as the most widely used media to commit these actions. Similarly, a recent survey on hate crimes in Spain 2021[2] shows that 41.65% of the participants, out of a total of 437, have been victims of hate crimes on more than one occasion in the last 5 years. On the one hand, they have received offensive comments on more than 10 occasions. On the other hand, more than 50% of them have received offenses or threats through social networks or the Internet. Finally, more than 70% of the respondents have received discriminatory treatment on one or more occasions in the last 5 years.

In this regard, inaction against offensive language allows for the further reinforcement of prejudices and stereotypes, while this type of hostile communication may lead to negative psychological effects among online users, causing anxiety, harassment, and, in extreme cases, suicide [3]. As a result, this scenario has motivated interested stakeholders (governments, online communities, and social media platforms) to look for efficient solutions to prevent Internet hostility. One strategy employed to tackle this problem is through legislation, by implementing laws and policies. For instance, since 2013 the Council of Europe has sponsored the "No Hate Speech" movement[3] seeking to mobilize young people to combat HS and promote human rights online. In May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter, and YouTube to implement the "Code of conduct on countering illegal HS online"[4]. From 2018 to 2020, platforms such as Instagram, Snapchat, and TikTok adopted the Code. One of the initial and most common approaches to hatred intervention adopted by social media platforms is content moderation. This approach is based on the suspension of

[1] `https://bit.ly/3xYhnZB`
[2] `https://bit.ly/3QjhrbX`
[3] `https://cutt.ly/sj5EdJ7`
[4] `https://bit.ly/2KI14cO`

user accounts and the removal of hate messages while attempting to balance the right to freedom of expression.

Although these approaches have the clear advantage of analyzing the context and accurately identifying this behavior, still these strategies do not seem to achieve the desired effect because they involve an intense, time-consuming, and costly procedure that limits scalability and quick solutions. At the same time, hate content is continuously growing and adapting, making it harder to identify [4]. As a result of these challenges, an alternative and preferable option is to rely on NLP-based methods to automatically detect this type of harmful online communication. Advances in NLP can be used to detect offensive content online and thus decreasing the time and effort in fighting this problem. Offensive language detection and analysis has become a major area of research in NLP. However, existing NLP-based methods face a number of drawbacks. Firstly, detecting offensive content is challenging for machines [5–7], since this type of language presents a subjective nature as well as social and cultural implications. Though recent approaches of sequence-to-sequence models [8, 9] have achieved good performance in detecting this type of content, most of them have not considered linguistic phenomena that may occur in the expression of offensive language such as those of an implicit nature such as sarcasm and irony [10, 11]. Secondly, since most of the available corpora contain messages from the Twitter platform, automatic systems have specialized in learning the language style and register used by the users on this platform, making cross-domain transfer difficult when employing such systems on other platforms. Thirdly, so far most of the research to solve this problem has been focused on English [12], leaving other languages such as Spanish in second place, despite the fact that combating this type of behavior is a global concern.

These challenges motivate this doctoral thesis to explore methods for accurately detecting offensive language on the Web using NLP techniques to aid in this process. **This thesis relies on advanced methods in NLP such as deep learning to tackle this issue**. First, it faces the problem of limited training data, especially in Spanish, generating appropriate resources to combat offensive textual content. These resources will also help to solve the limitation of the systems specialized in Twitter since messages from other social platforms such as YouTube and Instagram are considered. Secondly, it introduces different linguistic phenomena that could be involved in the expression of offensiveness and could help in the detection of this content. Then, a novel method is proposed where these identified phenomena are integrated for the detection of offensive language, using state-of-the-art techniques based on transfer learning. Finally, this novel method is applied for the detection of different offensive language scenarios (HS, sexism, toxicity), analyzing which specific linguistic phenomena are beneficial in each of them.

1.2 Research hypothesis

This thesis studies the problem of automatically detecting offensive textual language with deep learning techniques for NLP. The main hypothesis of this thesis is the following:

Main hypothesis

Advanced NLP methods based on Deep Learning, in particular Transfer Learning, aid in the detection of offensive textual language.

In particular, we subdivide this hypothesis into three hypotheses that will be addressed by the approaches proposed in this thesis:

Hypothesis 1 (H1). *The subjective nature of offensive language can have strong cultural, demographic, and social implications, and therefore language-specific resources and models are required.*

Language is a cultural carrier. It provides all cultural information, both verbally and in writing. Culture, on the other hand, influences and shapes language. The expression of offensiveness is one aspect of language on which culture has a great influence. Because every culture has a different concept of what is and is not a social norm, certain behaviors and language that are natural in one society are regarded as blasphemous and obscene in another. Therefore, we believe that in NLP research it is important to develop language-focused resources, rather than adapting resources from, for example, English, the language with the most developed resources. We examine the significance of building language-specific resources for offensive language identification by creating appropriate resources for Spanish (Section 4) and comparing the performance of multilingual and monolingual deep learning models for Spanish offensive language detection (Section 3).

Hypothesis 2 (H2). *Transfer learning models that leverage linguistic phenomena information related to the expression of offensive language, outperform models for offensive language detection that do not integrate this information.*

Many existing methods only consider the offensive language task as a sole optimization objective, however, the expression of offensive language implies both explicit and implicit phenomena that should be considered in NLP systems to better accurately this problem. To overcome this challenge, we propose a transfer learning method that integrates different linguistic phenomena to detect offensive language (Section 5). It relies on approaching different linguistic phenomena involved in the expression of offensive language as tasks in order to simultaneously learn via MTL learning common features among them and improve the generalization of the model.

Hypothesis 3 (H3). *Integrating specific linguistic phenomena into a transfer learning methodology can be beneficial in detecting different offensive scenarios. Offensive language detection comprises different scenarios, for instance, the identification of sexist content, HS detection, or the detection of toxic language.*

While these scenarios share some similarities, each of them also represents unique characteristics of the problem to be addressed. Therefore, we believe that the study of which phenomena are beneficial for each scenario should be considered while integrating them into a transfer learning methodology. To test this, we evaluate the integration of different linguistic phenomena in a variety of offensive language scenarios (Section 6).

1.3 Thesis outline

This thesis is divided into 8 chapters and organized as follows:

- **Chapter 2** includes an overview of the background information that is significant for understanding the content of this thesis. We review traditional ML and NN-based methods for offensive language research in NLP. We furthermore provide a compilation of different existing resources labeled with offensiveness. Then, we present the research challenges and opportunities based on the previous research approaches reviewed.

- **Chapter 3** introduces the preliminary research we conducted in the thesis, focusing mainly on traditional ML approaches to address HS detection, including misogyny and xenophobia. In addition, we present the first experiments with monolingual and multilingual pre-trained language models based on Transformers.

- **Chapter 4** describes the different corpora and lexicons we generate during the thesis for the research on offensive language and emotion analysis. Specifically, three corpora and three lexicons, mainly focused on Spanish, are presented.

- **Chapter 5** introduces our contribution to addressing offensive language detection. After an extensive review of previous methodologies, we propose a novel approach that uses the MTL paradigm to combine different phenomena that are inextricably related to the expression of offensive language. This approach aims to benefit from shared knowledge across tasks to improve the detection of offensive language. In this chapter, we define some of the linguistic phenomena that could be involved in the expression of offensive language and present the initial experiments with a subset of these phenomena (sentiments and emotions) on two Spanish corpora.

- **Chapter 6** focuses on the evaluation of the proposed MTL learning approach in different offensive language scenarios studying the integration of the linguistic phenomena defined in Chapter 5. The offensive scenarios tested in this chapter are the following: sexism identification in social networks, the detection of toxicity in comments, and HS and offensive content identification. We show the success of our MTL methodology by comparing its performance with previous state-of-the-art approaches that do not consider this useful information.

- **Chapter 7** presents two different shared tasks organized in the framework of this doctoral thesis to promote the research on emotion analysis and offensive language detection in Spanish. The task descriptions, the corpora and evaluation measures used as well as the participants and results achieved are described.

- **Chapter 8** finally summarizes our conclusions where we present the main findings of this doctoral thesis and suggest future research directions within offensive language research.

Chapter 2

Literature review

This chapter covers background information to set the stage for the following chapters. We define the offensive language problem and describe a taxonomy that comprises different concepts used in the literature. Then, we provide a compilation of linguistic resources most used for offensive language detection. After that, we give an overview of the NLP approaches applied to address the problem. Finally, we discuss research challenges and opportunities based on the previous work analyzed.

2.1 Offensive language

Offensive language is commonly referred to as derogatory, hurtful or obscene utterances [6], which may include insults, threats, profane language or swear words [13]. In the literature, closely related terms include HS, cyberbullying, toxicity, and profanity, among others. Offensive language is a complex phenomenon due to its highly subjective nature which may have strong cultural, demographic, and social implications, that is, an utterance could be considered offensive or not depending on one's cultural background. Furthermore, the language used to communicate this attitude can range from simple swear words or insults (1)-(3) to more difficult cases in which the offensive nature is conveyed by other means, such as sarcasm, mockery, and the use of negative stereotypes, among others (4)-(5).

> The examples throughout this doctoral thesis are included to illustrate the seriousness of the problem of offensive language. They in no way represent the perspective of the authors.

1. He is a **drunk crazy man**.

2. Getting scared **ugly man**?

3. The Liberals are **mentally unstable**!!

4. I have never had an intelligent conversation with a woman.

5. I would gladly invite Moroccans to go to their country.

The proliferation of offensive language in user-generated web content and, in particular on social media networks is steadily growing which makes it difficult or even impossible to manually track the content of comments. In recent years, interest in online offensive language detection and, specifically, in the automation of this attitude has continued to grow, along with the social impact of the phenomenon. The NLP area plays an important role as a powerful tool to automatically tackle this problem. Offensive language detection in NLP is commonly formulated as a binary or multi-class classification task [5, 14]. In the former, textual units are mapped to offensive or non-offensive classes (6)-(7) while in the latter different offensive categories are considered. For instance, automatic categorization of offense types (targeted insult, untargeted) (8), or offense target identification (individual, group, others) (9).

6. I can't stand the inept man who works at the bank. **Class: offensive**.

7. I love listening to music in the mornings. **Class: non-offensive**.

8. He is so **stupid** that no one talks to him. **Class: targeted insult**.

9. **Gay** pride day is overrated, it shouldn't exist. **Class: target to a group**.

2.2 Taxonomy of offensive language

The lack of consensus among researchers leaves open the possibility of subjective interpretations of offensive language identification problems. Several of these concepts found in the literature are abusive language [15], aggressive behavior [16], cyberbullying [17], HS [18], toxicity [19], and profanity [20]. In Table 2.1, the definition of these concepts are shown. While all the concepts presented are slightly different in meaning, they share common characteristics. Therefore, in order to group these concepts, different taxonomies have been proposed by researchers. For instance, Nobata et al. [15] distinguish between clean and abusive language, where the latter can be labeled as HS, derogatory or profane. Founta et al. [21] differs between HS and abusive/offensive language because HS involved a well-defined description of the target groups of this category, compared to the rest. Another taxonomy defined by Poletto et al. [7] considered the offensiveness

Term	**Definition**	**Source**
Abusive Language	"Any strongly impolite, rude, or hurtful language using profanity and also HS".	Nobata et al. [15]
Aggressive Behavior	"Overt, angry and often violent social interaction, conducted through online media, with the intent to inflict harm or discomfort on another individual or group of people, who perceive such acts as derogatory, harmful or unwanted."	Chatzakou et al. [16]
Cyberbullying	"Aggressive and intentional acts conducted by an individual or group or, using online media, repeatedly and over time, against a victim who is unable to easily defend himself/herself."	Chen et al. [17]
Hate Speech	"The advocacy, promotion, or incitement, in any form, of the denigration, hatred, or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization, or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of race, color, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation, and other personal characteristics or status."	European Commission [18]
Toxicity	"When it attacks, insults, offends, or disqualifies a person or group of people on the basis of characteristics such as race, ethnicity, nationality, political ideology, religion, gender, and sexual orientation, among others."	Taulé et al. [19]
Profanity	"Offensive or obscene word or phrase."	Cambridge Dictionary [20]

TABLE 2.1: Related concepts to Offensive Language.

and HS as instances of abusive language, whereas [22] considered the offensive language as an umbrella term for all related concepts (abusive language, HS, toxicity, or aggressiveness). Since the problems studied in this thesis are all compatible with the general definition of offensive language, we decided to adopt the taxonomy of Kogilavani et al. [22] shown in Figure 2.1 along with the research conducted.

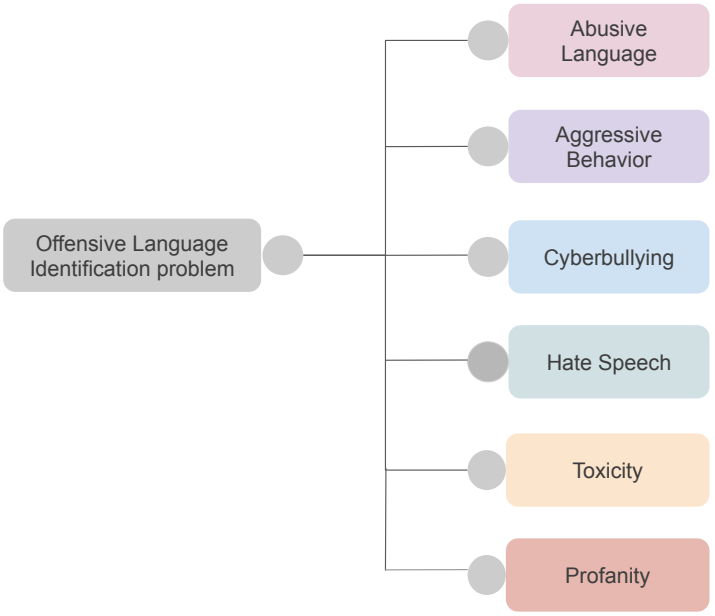

FIGURE 2.1: Taxonomy of offensive language identification problem. This is the taxonomy adopted in this doctoral thesis.

2.3 Shared task evaluation campaigns

In recent years, the interest in tackling this important problem by the NLP community has been reflected in the number of shared tasks proposed to encourage offensive language research. The organization and participation in these campaigns are very important in order to make progress in the task. The following is a review of the most relevant campaigns in the area of offensive language.

One of the first was **GermEval** [23] on the identification of offensive language which was organized in 2018 and focused on classifying German tweets. Two tasks were proposed, a coarse-grained binary classification that aims to decide whether a tweet includes some form of offensive language or not, and a fine-grained 4-way classification that involves four categories (*profanity*, *insult*, and *abuse*, *other*). Overall, 20 teams participated in the shared task. The same formulation was continued in the 2019 edition [24].

The Task on Automatic Misogyny Identification (**AMI**) [25] was held in the workshop Evaluation of NLP and Speech Tools for Italian (Evalita) in 2018. It focused on both Italian and English tweets and two subtasks were proposed. The first subtask deals with misogyny identification and the second subtask address misogynistic behavior categorization and target classification. A total of 16 teams participated in submitting their systems.

In 2019, different shared tasks were proposed. the Hate Speech and Offensive Content Identification in Indo-European Languages (**HASOC**) [26] created datasets for Hindi,

German, and English for the identification of HS and offensive language on Twitter and Facebook posts. HASOC continued in 2020 by introducing two tasks, one on coarse-grained HS and offensive language vs. non-HS non-offensive language and one which distinguishes hate, offensive language, and profane language for all these languages. HASOC 2021 was extended by a subtask on code-mixed language.

Another shared task organized in 2019 was **HatEval** [27] on HS detection against immigrants and women was held as part of the International Workshop on Semantic Evaluation (SemEval). It was focused on Spanish and English tweets. Two classification subtasks were organized: a binary subtask to detect the presence of HS, and a fine-grained one aimed at identifying other features in hate content, including aggressive attitude and harassed target.

As part of the same workshop SemEval, another popular shared task named **OffensEval** was proposed, which was held in two editions, 2019 and 2020 [8, 13]. In both editions, they proposed a three-level taxonomy to address three different subtasks: (a) offensive language identification, (b) automatic categorization of offense types, and (c) offense target identification. In the first edition, the Offensive Language Identification Dataset (OLID) was released which contains over 14,000 English tweets and a total of 115 teams submitted their participation. In the second edition, the same three-level taxonomy was used to release a multilingual dataset in five languages: English, Turkish, Arabic, Danish, and Greek. This edition received a total of 70 system description papers.

A recently shared task that took place in 2021 and was held as part of the Iberian Languages Evaluation Forum (**IberLEF**) is **MeOffendEs** [28] which is one of the tasks organized within the framework of this doctoral thesis (see Chapter 7: *"Organized Shared Tasks"*). This shared task event was focused on the identification of offensive language for Spanish variants and involved four subtasks: the first two correspond to the identification of offensive language categories in generic Spanish texts from Twitter, Youtube, and Instagram, while subtasks 3 and 4 were related to the identification of offensive language targeting the Mexican variant of Spanish in Twitter. The same year, two different shared tasks in IberLEF were proposed. **EXIST** focused on sexism identification in social networks [29] proposed two challenges both in Spanish and English: sexism identification and sexism categorization of tweets and gabs. Finally, **DETOXIS** [19] suggested the challenge of detecting toxicity in comments posted in Spanish in response to online news articles related to immigration. Two subtasks were structured: a binary classification task to detect toxicity and a multiclass classification task to detect the level of toxicity. A total of 31 teams participated in this shared task.

These shared tasks motivated the creation of linguistic resources for different languages such as English, German, Italian, Spanish, Hindi, and others. However, beyond these

shared tasks most of the research and language models are still mainly focused on English and therefore considerable efforts should continue to be achieved for other languages such as Spanish where the number of resources remains low.

2.4 Linguistic resources for offensive language detection

In this section, a compilation of both corpora and lexicons in the frame of offensive language detection[1] is going to be presented. Please note that this compilation does not cover all the resources developed, but the most relevant in the area.

2.4.1 Corpora

The corpora annotated with offensiveness most used in the literature are going to be described. Most of these corpora have been generated within the framework of shared tasks, some of them described in the previous section. To the best of our knowledge, there are corpora annotated for English, Spanish, Italian, German, Hindi, Tamil, Greek, Turkish, Danish, Arabic, Marathi, and Malayalam.

We are going to review corpora in different languages, both in English, the language with the most resources in this field, and in other languages where resources have also been created, with a special focus on Spanish, which is the main language object of study in this doctoral thesis. They are summarized in Table 2.2. We consider this compilation may be valuable for the scientific community to advance in the study of this phenomenon in different languages.

HS_OFF_EN

One of the first datasets to appear in the context of offensive language was built by Davidson et al. [4] in 2017. We refer to this dataset as HS_OFF_EN. Authors used the Twitter API to search for English tweets containing terms from the `Hatebase.org`[2] lexicon, resulting in a sample of tweets from 33,458 Twitter users. Then, the timeline for each user is extracted, resulting in a set of 85.4 million tweets. From this corpus, they selected a random sample of 25,000 tweets containing terms from the lexicon. After that, the annotation took place on the CrowdFlower platform. Workers were asked to label each tweet as one of three categories: HS, offensive but not HS, or neither offensive nor HS. The intercoder-agreement score was high, 92%. Authors use the majority decision

[1] We consider offensiveness as an umbrella to encompass different terms employed in the literature such as HS, aggressiveness, toxicity, and sexism, among others.

[2] https://hatebase.org/

for each tweet to assign the final label. The final dataset contains a total of 25,000 English tweets.

TRAC

The TRAC dataset was provided for the first time by the organizers of the Shared Task on Aggression Identification as part of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) in 2018 [30]. This corpus contains 15,000 aggression-annotated Facebook posts and comments in Hindi and English. Each post is annotated with 3 levels of aggression - overtly *aggressive*, *covertly aggressive* and *non-aggressive*. This dataset is a subset of a larger dataset discussed in [31]. In 2020, the second edition of the workshop, the organizers provided another dataset including the Bengali language. It contains approximately 6,000 sampled Youtube comments for Bengali, Hindi, and English. They included another level to annotate instances as *gendered* or *non-gendered*.

HS_IT

In 2018, Sanguinetti et al. [32] built one of the first Italian corpora for offensive language detection composed of 6,000 tweets annotated for HS against immigrants. We refer to this corpus as HS_IT. The collection contains tweets gathered using a traditional keyword-based technique, specifically by filtering the corpus using neutral keywords associated with three social groups identified as potential HS targets in the Italian context: immigrants, Muslims, and Roma. Following a first annotation step that yielded a collection of approximately 1,800 tweets, the corpus was enlarged by adding new annotated data. The new tweets were annotated by professionals as well as CrowdFlower contributors. This corpus is annotated with the following categories: *HS*, *aggressiveness*, *offensiveness*, *irony*, *stereotype*, and *intensity*.

OLID

The OLID dataset was provided by the organizers of SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) [13]. OLID is a large collection of 14,100 English tweets that have been labeled using a three-layer hierarchical annotation methodology (whether the tweet is offensive or not, the type of the offense, and the target to whom the offense is directed). The dataset annotation was placed on Figure Eight crowdsourcing platform[3]. Authors ensured annotation quality by only hiring experienced annotators on the platform and utilizing test questions to exclude annotators who did not meet a specified threshold. Two people annotated all of the tweets. In the case of a disagreement, a third annotation was requested, and finally, a majority vote was used.

[3] `https://appen.com/`

HASOC

The series of HASOC datasets started in 2019 with the Hate Speech and Offensive Content Identification in Indo-European Languages shared task at FIRE [26]. This dataset was subsequently sampled from Twitter and Facebook for different languages (English, Hindi, and German). In order to retrieve these tweets, authors used hashtags and keywords that contained offensive content. Several students from each language used an online system to judge the tweets during the labeling procedure. Each tweet was annotated in different levels: Level A (*hate and offensive*, *non hate-offensive*), Level B (*hate*, *offensive*, *profane*), Level C (*targeted insult*, *untargeted*). After calculating the inter-annotator agreement, the authors found that English was the language with the least agreement, followed by German and Hindi. For the 2020 edition of the shared task, one of the main objectives of the authors was to minimize the impact of bias in the data offered in 2019. Therefore, they develop an HS dataset based on a sampling process that relies on less input. They used an available tweet collection named archive.org[4] and they downloaded the entire archive corresponding to May 2019. After downloading the archive of tweets, they identified English, German, and Hindi tweets using the language attribute provided by the Twitter metadata. A selection of tweets was annotated manually by people who use social media in their respective languages. They followed the same levels of annotation as in 2019. For the annotator agreement, in this case, they obtained the lease agreement for the Hindi language, followed by English and German. For the third edition of HASOC 2021 [33], authors incorporated tweets for Indo-Aryan languages including Hindi, and Marathi. The dataset collection was performed when India was facing the second and extremely hard COVID-19 wave.

HatEval

The HatEval dataset was provided by the organizers of the SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [27]. The data were collected using different gathering strategies for both English and Spanish. For what concerns the time frame, tweets were mainly collected in the time span from July to September 2018, with the exception of data with target women. Indeed, most of the training set of tweets against women was derived from an earlier collection carried out in the context of two previous challenges on misogyny identification, whose collection phase started on the 20th of July 2017 and ended on the 30th of November 2017. The data were released after the annotation process, using the crowdsourcing platform Figure Eight (F8). They were required to collect at least three independent judgments for each tweet. They adopted the default F8 settings for assigning the majority label (relative majority). The authors assigned the final label for this data based on majority

[4]https://archive.org/details/archiveteam-twitter-stream-2019-05

voting from the annotators. The final dataset is composed of over 6,000 tweets in each language.

HaterNet

HaterNet is a Spanish dataset that was collected with an intelligent system used by the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security [34]. The first step in the creation of this corpus was to collect tweets on different random dates between February 2017 and December 2017. A final collection of 2 million tweets originating from Spain was retrieved. The second step was to apply a filter prior to the manual labeling process. The filter was generated using six dictionaries of HS and one dictionary that contains generic insults. The elements of HS dictionaries were labeled with one of two possible degrees of hate: *absolute* or *relative*. If the tweet contained at least one absolute element of these HS dictionaries, then it was selected as a possible container of HS. If, on the other hand, it contained at least one relative element of these HS dictionaries and at least one element of the swearword dictionary, it was also selected as a likely container of HS. After that, only 8,710 tweets were selected for manually labeling. The third step was the labeling of the selected tweets by four experts with different backgrounds: a 44-year-old public servant, a 23-year-old graduate in Psychology, a 24-year-old Law graduate, and a 23-year-old Criminology graduate. The final label of each tweet was decided by a majority vote, and in the case of a tie a fifth person, a 49-year-old professor of Computer Science, cast the deciding vote. Finally, the dataset is composed of 6,000 tweets, The average inter-agreement among the labelers using Fleiss' kappa [35] was 0.588.

SOLID

The Semi-Supervised Offensive Language Identification Dataset (SOLID) is an extension of the OLID dataset used for the shared task SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media [8]. The authors collected random tweets utilizing the 20 most common English stopwords such as *the*, *of*, *and*, *to*, etc. The collected tweets were then labeled semi-supervised using democratic co-training and OLID as a seed dataset. For the co-training, they used four ML models with different inductive biases: PMI [36], FastText [37], LSTM [38], and BERT [39]. They used this semi-supervised technique to choose the offensive tweets for the test set, and then manually tagged them for the various categories. The SOLID dataset contains 9,089,140 English tweets, which makes it the largest dataset of its kind.

ALT

The Arabic offensive dataset was provided by the organizers of the SemEval-2020 Task 12: Arabic and English offensive language identification in social media [40]. It consists

of 10,000 tweets collected between April and May of 2019 using the Twitter API with the language filter set to Arabic. Only tweets with two or more vocative particles were considered for annotation in order to increase the likelihood of offensive content; the vocative particle is used primarily to direct speech to a person or group, and it is commonly observed in offensive communications in almost all Arabic dialects. This resulted in 20% offensive tweets in the final dataset. A native speaker familiar with many Arabic dialects manually annotated tweets as *offensive* or *non-offensive*. A random subsample of offensive and non-offensive tweets was double annotated, and the Fleiss *kappa* obtained was 0.92. This dataset was also used in the SemEval-2020 OffensEval competition [8].

DKhate

The Danish dataset, DKhate, [41] is composed of 3,600 user-generated comments retrieved from Facebook, Reddit, and a local newspaper, Ekstra Bladet3[5]. The selection of comments was partially seeded with offensive language acquired during a crowd-sourced lexicon compilation. This dataset is annotated for various types and targets of offensive language. The annotation was performed at the individual comment level by males aged 25-40. This dataset was used in the SemEval-2020 OffensEval competition [8].

OGTD

The Offensive Greek Twitter Dataset (OGTD) [42] includes 10,287 tweets sampled using popular and trending hashtags, including television shows such as series, reality, and entertainment shows, as well as some politically related tweets. Another portion of the dataset was retrieved using pejorative terms and the keywords "you are" This technique was chosen with the expectation that TV and politics would gather a small number of offensive posts, as well as tweets containing vulgar language, for further study. A team of three volunteer annotators participated in the annotation process. Each tweet was annotated as *offensive* or *non-offensive*. In cases of disagreement, labels with a majority agreement above 66% were selected as the final tweet labels. This dataset was used in the SemEval-2020 OffensEval competition [8].

Turkish_OFF

The Turkish corpus is the first dataset annotated with offensive language for this language [43]. It contains 36,232 messages selected at random from the Twitter stream between April 2018 and September 2019. The annotators were volunteers recruited from the author's contacts. All of the annotators are native speakers of Turkish, and all

[5] `http://ekstrabladet.dk/`

are highly educated. Cohen's κ calculated on 5,000 doubly-annotated tweets was 0.761. Authors found that around 19% of the tweets in the data contain some form of offensive language, which are further classified based on the target of the offense. This dataset was used in the SemEval-2020 OffensEval competition [8].

EXIST

The series of EXIST datasets started in 2021 with the sEXism Identification in Social neTworks shared task [29, 44] which has been held for two editions at IberLEF 2021 and 2022. This multilingual dataset in English and Spanish incorporates any type of sexist expression or related phenomena, including descriptive or reported assertions where the sexist message is a report or a description of sexist behavior. Popular expressions and terms, such as those used in previous approaches to the state of the art, both in English and Spanish, used to undervalue the role of women have been extracted from various Twitter accounts and analyzed and filtered by two gender experts. The final set contains more than 200 expressions that can be used in gendered contexts. Using the final set of sexism terms (94 seeds for Spanish and 91 seeds for English), tweets were extracted in both languages (over 800,000 tweets were downloaded). Final labels of tweets were selected according to the majority vote between five crowdsourcing annotators, who followed the guidelines developed by the experts, but tweets with 3 to 2 votes were manually reviewed by two people with more than two years of experience analyzing sexist content in social networks. As a result, the multilingual dataset in the first edition had over 11,345 instances from Gab and Twitter. For EXIST 2022 challenge, the authors continue organizing the shared task with the same formulation of tasks and provide the dataset with more instances.

NewsCom-TOX

The NewsCom-TOX corpus was provided by the organizers of the DETOXIS shared task [19] that was held as part of IberLEF 2021. This corpus contains comments with toxic language. Specifically, the corpus consists of 4,359 comments posted in response to 21 different articles extracted from Spanish online newspapers (ABC, elDiario.es, El Mundo, NIUS, etc.) and discussion forums (such as Meneame and ForoCoches) from August 2017 to July 2020. These articles were manually selected taking into account their controversial subject matter, their potential toxicity, and the number of published comments (minimum of 50 comments). A keyword-based approach was used to search for articles predominantly related to immigration. The number of comments ranged from 65 to 359 comments per article. On average, 31.16% of comments are toxic. In addition, each post is labeled with different features including *target group*, *constructiveness*, *mockery*, *sarcasm*, *target person*, *insult*, *stereotype*, *improper language*, *intolerance*, *aggressiveness*, *argumentation*, *positive stance*, and *negative stance*.

Dataset	Task	Source	Language	Size	Reference
HS_OFF_EN	HOF	Twitter	EN	25,000	Davidson et al. 2017
TRAC	Aggressiveness	Facebook	EN, HI	15,000	Kumar et al. 2018
HS_IT	Xenophobia	Twitter	IT	6,000	Sanguinetti et al. 2018
OLID	OFF	Twitter	EN	14,100	Zampieri et al. 2019
HASOC	HOF	Twitter, Facebook	EN, GER, HI	17,657	Mandl et al. 2019
HatEval	Misogyny and Xenophobia	Twitter	ES	$\sim$ 6,600 per lang.	Basile et al. 2019
HaterNet	HS	Twitter	ES	6,000	Pereira-Kohatsu et al. 2019
SOLID	OFF	Twitter	EN	$\sim$ 9 millions	Zampieri et al. 2020
ALT	OFF	Twitter	ARA	10,000	Hassan et al. 2020
DKhate	OFF	Facebook, Reddit, newspapers	DA	3,600	Sigurbergsson and Derczynski 2020
OGTD	OFF	TV shows	GRE	10,287	Pitenis et al. 2020
Turkis_OFF	OFF	Twitter	TR	10,287	Çöltekin 2020
EXIST	Sexism	Twitter, Gab	EN, ES	11,345	Ródriguez-Sánchez et al. 2021
NewsCom-TOX	Toxicity	Newspapers	ES	4,357	Taulé et al. 2021

Table 2.2: Dataset most commonly used in recent years for the detection of tasks in the context of offensive language. HOF: Hate Speech and Offensive Language, OFF: Offensive Language, EN: English, ES: Spanish, DA: Danish, GER: German, HI: Hindi, ARA: Arabic, TA: Tamil, MAR: Marathi, ML: Malayalam, TR: Turkish, GRE: Greek, lang.: language

2.4.2 Lexicons

In this section, some of the lexicons developed by the NLP community and annotated with any definition which involved offensive language are going to be described.

A lexicon of abusive words

Wiegand et al. [45] were one of the first in the NLP community to create abusive lexicons in English, specifically they built a base lexicon and an expanded lexicon. The base lexicon is a small set where the terms were obtained from the Subjectivity Lexicon [46] which contains negative polar expressions, specifically they sampled 500 negative nouns, verbs, and adjectives. In addition, the authors added some prototypical abusive words missed in this lexicon such as "nigger", "slut", or "cunt". These terms were annotated via crowdsourcing by 5 native English annotators. The base lexicon follows a binary word categorization: *abusive* and *non-abusive*. A word was labeled as abusive if at least four of the five annotators judged it as abusive. The authors decided to expand this lexicon by categorizing all (unlabeled) negative polar expressions from Wiktionary. The negative polar expressions are identified by using an SVM trained on words from the Subjectivity Lexicon with their corresponding polarity to the Wiktionary vocabulary. They used word embeddings as features. Another SVM was trained on the base lexicon to generate the feature-based lexicon of abusive terms. Finally, this lexicon has 2,989 offensive terms, which is 5 times the size of the base lexicon.

HurtLex

Another popular resource is HurtLex [47], a multilingual lexicon of hate words that covers over 50 languages and is organized into 17 categories such as derogatory words, physical disabilities and diversity, negative stereotypes, and ethnic slurs. Authors started from a preexisting Italian lexical resource [48] to perform a semi-automatic multilingual extension using MultiWordNet [49] and BabelNet [50]. Hurtlex contains two levels of structure: conservative, which is obtained by translating offensive senses of words in the original lexicon, and inclusive, which is obtained by translating all potentially relevant senses of words in the original lexicon.

Hatebase

Hatebase[6] is a collaborative repository of multilingual HS which contains terms related to the expression of HS. It has been developed to assist companies, government agencies, NGOs, and research organizations to moderate online conversations. It comprises a broad multilingual vocabulary based on nationality, ethnicity, religion, gender, sexual discrimination, disability, and class to monitor incidents of hate speech across countries,

[6] `https://hatebase.org/`

specifically it is composed of 3,894 terms, 98 languages, and 175 countries. For Spanish, 142 terms can be found.

2.5 NLP approaches for offensive language detection

In this section, an overview of the different approaches that have been applied in recent years to perform the offensive language detection task is provided, from early approaches based on traditional algorithms to more recent techniques focused on transfer learning. Furthermore, we will review a variety of studies that take into account related phenomena to offensive language (emotions, sentiments, etc.) to combat this problem. First, each of these approaches will be explained in detail, then the studies that have used them for the detection of offensive language will be mentioned.

2.5.1 Traditional approaches

Among the most common ML approaches employed in the literature, we can differentiate three different categories: supervised learning, unsupervised learning, and semi-supervised learning. In supervised learning, a model is designed to train a classifier that requires annotated data to learn the specific patterns for the task. Once the model is trained, it is used to predict new instances which are not labeled. In contrast, unsupervised methods do not rely on annotated data to learn about the task but use, for example, external resources such as lexicons to define a heuristic and address the task. Semi-supervised learning combines these two methodologies defined to achieve the benefit of multiple methods and reach the maximum level of accuracy.

Most studies on offensive language identification have adopted supervised learning methods because of their success in obtaining good performance in several NLP tasks, however, they require a large number of labeled resources to be trained. As part of this family, we found statistical machine learning approaches and NN which are going to be described in detail below.

2.5.1.1 Statistical Machine learning methods

Initially, traditional statistics models such as Naive Bayes, Support Vector Machines, Logistic Regression, and Decision Trees were among the most popular methods to address this task [12, 51]. In the following, these classifiers are described in detail.

Naive Bayes (NB). Naive Bayes is a probabilistic classifier method based on Bayes' theorem [52]. Naive Bayes has been successfully applied to document classification in many research efforts [53]. In this study, the Multinomial Naive Bayes classification model was used. This model is suitable for classification with discrete features like word frequency information in a document, where a document is a sequence of words obtained from vocabulary 'V'. The probability of a document given its class can be obtained using the multinomial distribution shown in Equation 2.1:

$$P(d_i|c_j;\theta) = P(|d_i|)|d_i|!\prod_{t=1}^{|V|}\frac{P(t|c_j;\theta)^{N_{it}}}{N_{it}!} \tag{2.1}$$

where $P(d_i|c_j;\theta)$ is the probability of document 'd' for each class 'c'. $P(|d_i|)$ is the probability of document 'd' and $P(t|c_j;\theta)$ is the probability of occurrence of a term 't' in a class 'c'.

There are other varieties of NB classifiers, such as multinomial Naive Bayes, which are commonly used for document classification tasks, Bernoulli Naive Bayes with boolean variables as predictors, and Gaussian Naive Bayes when the predictors are continuous and not discrete.

Support Vector Machine (SVM). SVM is a linear learning technique that finds an optimal hyper-plane to separate our two classes (hateful and not hateful speech). Many researchers have reported that this classifier is perhaps the most accurate method for text classification [54] and also is widely used in sentiment analysis [55]. In this paper, linear SVM is used. The formula for the output of a linear SVM can be represented as:

$$u = \vec{w}^{\,t} \cdot \vec{x}^{\,t} - b \tag{2.2}$$

where $\vec{w}^{\,t}$ is the normal vector to the hyperplane, and $\vec{x}^{\,t}$ is the input vector.

The SVM model allows the expansion of space through kernels [56]. There are various kernels, the most common of which are linear, polynomial, Gaussian Radial Basis Function (RBF), and hyperbolic tangent or sigmoid.

Logistic Regression (LR). Logistic regression is a statistical method for predicting binary classes. Specifically, the algorithm LR is a discriminative model that describes the conditional probability as:

$$P(y|X) = \frac{exp(\sum_{m=1}^{M}\lambda_m f_m(y,X))}{\sum_{y'} exp(\sum_{m=1}^{M}\lambda_m f_m(y',X))} \tag{2.3}$$

In order to optimize the parameters of LR in our experiment, we used the solver parameter equal to liblinear.

Decision Tree (DT). A decision tree algorithm is a flowchart-like tree structure where an internal node represents features, the branch represents a decision rule, and each leaf node represents the outcome. In the context of text data, tree internal nodes are labeled by terms, branches are labeled by testing the weight, and leaf nodes are represented by the corresponding class. The tree can classify the document by running through the structure from the root until it reaches a certain leaf, which represents the goal for the classification of the document.

2.5.1.2 Statistical Machine learning methods for offensive language detection

Some of the initial studies that adopted these methods for addressing the offensive language detection task are the following. Chen et al. [17] presented a Lexical Syntactic Feature (LSF) architecture for detecting offensive content and identifying potentially offensive people on social media. For this aim, they incorporate different features related to the user into two different classifiers: SVM and NB, being the SVM the most successful classifier to predict this content, with a precision and recall of 0.78. Davidson et al. [4] trained a variety of models namely LR, NB, DT, RF, and linear SVMs to distinguish between *HS*, *offensive*, and *neither* categories. They found that LR and Linear SVM models tended to perform significantly better than others, with the best performing model obtaining an overall F_1-score of 0.90. Also, based on their results, they indicated that fine-grained labeling can aid in HS identification and highlight some of the challenges to accurate categorization. Pamungkas et al. [57] proposed an SVM-based architecture for misogyny detection in English and Spanish and explored the use of several sets of features, including a wide range of lexical features relying on the use of available and novel lexicons of abusive words (Hurtlex) [47]. They achieved a 0.91 in English and a 0.82 in Spanish in terms of accuracy. Finally, Malmasi and Zampieri [58] developed an SVM classifier combining different features (n-grams, skip-grams, and clustering-based word representations) to address the difficulty of separating ordinary profanity from HS in social media and achieved an accuracy of 0.80. The analysis of the results by the authors indicated that discriminating HS and profanity is a complex task that may require linguistic features that capture a deeper understanding of the context. Most of these studies have shown that one of the best-performing traditional ML classifiers for offensive language detection is the SVM. In addition, they show that the integration of linguistic features in these types of models is useful to help in the detection of offensive language. Some of the advantages and disadvantages of using these classifier methods

are the following: they do not need a large training set in order to achieve good results, and in addition, they are easy to interpret. However, they are not flexible enough to capture more complex relationships naturally.

2.5.1.3 Neural Networks

Another series of traditional methods for the detection of offensiveness are the NN models. Deep NN are a part of the Artificial NN (ANN) family of ML technologies. They are computing systems that are inspired by the organic NN that constitutes human brains. In the 2000s, NN started being used in NLP tasks. In 2003, the first neural language model, consisting of a single layer feed-forward NN, was proposed by Bengio et al. [59]. They were among the first to use dense feature vectors instead of sparse high-dimensional vectors to represent words, which became known as *word embeddings*. A set of vectors of word embeddings is the representation of the ideal semantic space of words in a real-valued continuous vector space, hence the relationships between vectors of words mirror the linguistic relationships of the words. Vectors of word embeddings are a dense representation of the meaning of a word, thus each word is linked to a real-valued continuous vector of dimension d_{emb}. In 2008, Collobert and Weston [60] showed that a unified NN could learn different NLP tasks while avoiding hand-crafted features and prior knowledge of the tasks. In 2013, Mikolov et al. [61] introduced one of the most popular word embedding models. Although using feature vectors to represent words was not a novel notion, they could speed up learning by simplifying the model and training it on vast volumes of textual data. Further studies [61, 62] showed that using pre-trained vectors to initialize NN that use word embedding as feature vectors increase the models' performance in several NLP tasks. The years 2013 and 2014 are considered to be the start of the widespread use of NN in NLP. Among them, Recurrent Neural Networks (RNN) [63] gained more popularity since they could better process sequences of different lengths in NLP and capture the context. However, vanilla RNNs were quickly replaced by **Long-Short Term Memories** (LSTM) [38] to address the longer dependencies between the words [64]. These networks operate at the word level and each sentence is represented as a sequence of word representations that are sequentially fed to the model one after another until the sequence has been entirely used up. These networks are trained using backpropagation through time and have memory blocks capable of learning temporal sequences and their long-term dependencies. A typical LSTM network is comprised of different memory blocks (see Figure 2.2) called cells (the rectangles in the image). Two states are being transferred to the next cell; the cell state and the hidden state. Memory blocks are in charge of remembering information, and modifications of this memory are carried out via three major mechanisms known as gates.

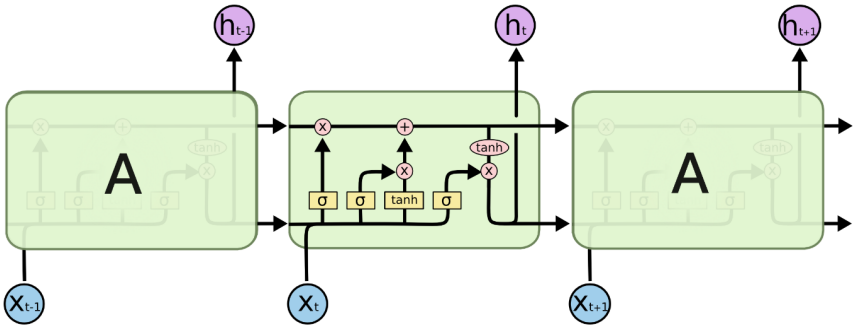

FIGURE 2.2: Long Short-Term Memory cell.

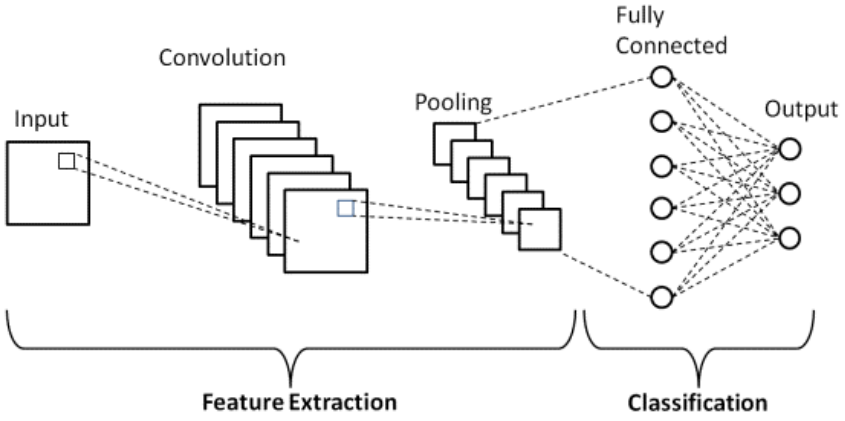

FIGURE 2.3: Architecture of a basic CNN.

Other types of NN are **Convolutional Neural Networks** (CNNs), which could parallelize the computation with the cost of capturing only the local context instead of taking the global representation [65]. This method was originally developed to be used in computer vision, but it has been shown to be effective for NLP tasks and specifically has achieved accurate results in text classification [66]. A CNN is composed of three different layers: convolutional layers, pooling layers, and fully-connected layers. In Figure 2.3, the general architecture of a basic CNN is shown. The first layer used to extract the various features from the input is convolutional. This layer performs the mathematical operation of convolution between the input and a filter of size MxM. This layer is usually followed by the pooled layer which aims to decrease the size of the map of convolutional features to reduce computing costs. These two layers are used for feature extraction. Finally, the fully connected layer which includes the weights and biases as well as the neurons is utilized to connect the neurons between layers. These layers are often placed prior to the output layer and constitute the last layers of the CNN architecture, and are in charge of inference for classification.

2.5.1.4 Neural Networks for offensive language detection

Both LSTMs and CNNs were among the most popular architectures adopted for offensive language detection. In the following, some of the studies that used these NN architectures are going to be described. For instance, Gambäck and Sikdar [67] developed a system for Twitter HS text identification based on two CNNs and feature embeddings including one-hot encoded character n-gram vectors and word embeddings, and they reported that the use of character n-gram does not help in the detection. In order to break the barrier of language dependency in the word embedding approach, Pitsilis et al. [68] conducted an ensemble of RNN classifiers, incorporating various features associated with user-related information. Paetzold et al. [69] experimented with a robust system based on compositional RNNs able to handle even substantially noisy inputs and reached competitive results for HS detection in English texts. Goenaga et al. [70] employed a BI-LSTM with Conditional Random Fields (CRF) in order to prove its effectiveness in misogynous tweet identification, obtaining 78.9 of accuracy on English tweets and 76.8 on Spanish tweets. Authors mentioned that identifying misogynous content in Spanish tweets is more difficult, owing to a lack of high-quality resources in comparison to English. Ribeiro and Silva [71] classified HS against women and immigrants in a multilingual context (English and Spanish) employing a CNN network using as word embeddings the GloVe vocabulary computed from the Spanish Billion Word Corpus (SBWC) and fastText from the Spanish Wikipedia, achieving a better performance in Spanish than English (69.6 and 48.8 F_1-score, respectively) and hypothesized that the reason could be in the nature of the Spanish corpus since it contains fewer tweets and there is a lack of the presence of complex phenomena like sarcasm or irony. Zampieri et al. [5] used both BiLSTM and CNN to predict the type and target of offensive posts in tweets, achieving the best performance with the CNN architecture. Corazza et al. [72] proposed a robust neural classifier for the HS classification task across different languages (English, Italian and German), and studied the impact of using different linguistic features and components (type of embeddings, the use of additional features - text-based or emotion-based - the role of hashtag normalization, and emojis) on the results across these languages. As NN architectures, they used LSTM, BiLSTM, and Gated Recurrent Unit. Authors found that (1) using subword information benefits the task because it allows to deal with social media's great language variety and creativity domain, as well as typos, (2) creating customized embeddings that cover the topic of interest well is advantageous to task performance, (3) hashtag normalization is beneficial for categorizing HS in English and Italian, however, it may not perform well on languages with high compound density, such as German, and (4) given the restricted length of tweets, LSTM outperformed BiLSTM. Finally, we could observe from these studies that traditional NN work well for the problem of offensive language detection

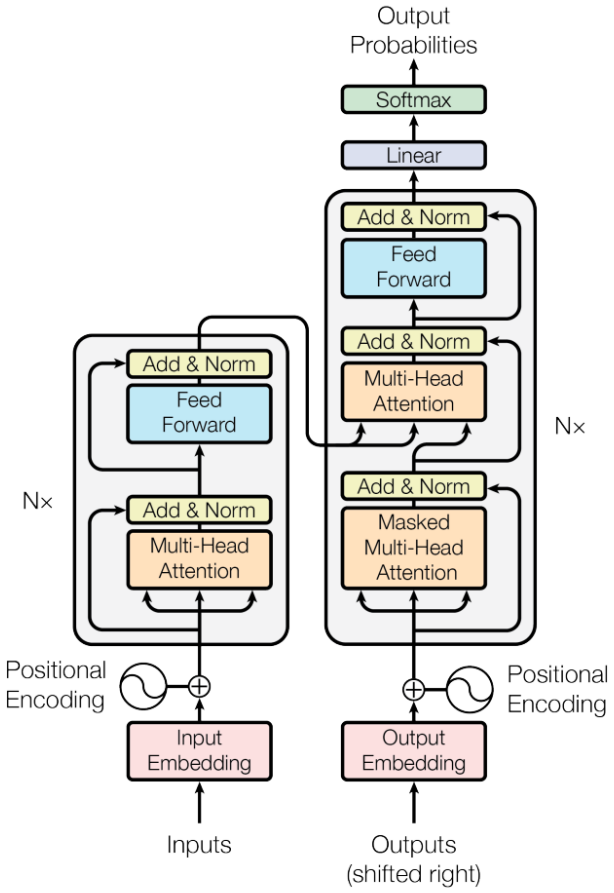

FIGURE 2.4: The arquitecture of the Transformer (Figure source: Vaswani et al. 2017).

but the results will vary depending mainly on the availability of linguistic resources used to train the model (both corpora, word embeddings, and lexicons), and the language. These methods also come with some disadvantages. For instance, they need massive amounts of available data in the training phase to achieve good performance, which is not always available due to the time-consuming process.

2.5.2 Transformer-based models

The Transformer's ground-breaking architecture was introduced in 2017 [1] and is currently state-of-the-art in several NLP tasks. The most innovative idea behind this architecture is the attention mechanism which allows learning contextual relations between words (or sub-words) in a text and includes two separate mechanisms: an encoder that reads the text input and a decoder that produces a prediction for the task [1]. As opposed to sequential models like RNNs, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. This characteristic allows the model to learn the context of a word based

on all of its surroundings and allows for significantly more parallelization than RNNs, resulting in shorter training durations. In this architecture, the encoder and decoder consist of N similar stacked layers. Each layer of the encoder consists of a self-attention and a position-wise feed-forward sub-layer. The decoder layers have an extra attention sub-layer, which also attends to encoder representations. The one-layer Transformer architecture is depicted in Figure 2.4. In the following, we describe each part of this architecture.

Self-Attention

The Transformer architecture is based on a novel concept inspired by the attention mechanism. By computing the representation of each position in the sentence directly from the last layer representations, the self-attention layer avoids the recurrence function. As a result, the representation of each position can be computed in parallel with other positions in the same layer. Each position's representation is computed in the self-attention layer by attending to all of the positions in the sequence. Unlike RNNs, where attendance to closer positions was stronger, the model in the self-attention layer can decide the relevance of the other positions to compute the representation. More formally, first, each input representation $\mathbf{x}_i$ at position i-th from the last layer is projected to three different vectors: key $\mathbf{k}_i$, query $\mathbf{q}_i$, and value $\mathbf{v}_i$. This is done by multiplying $\mathbf{x}_i$ by the projection matrices $\mathbf{W}_k \in \mathbb{R}^{d_m \times d_k}$, $\mathbf{W}_q \in \mathbb{R}^{d_m \times d_k}$, and $\mathbf{W}_v \in \mathbb{R}^{d_m \times d_v}$, where d_m is the model's hidden size, d_k is the key and query vector sizes, and d_v is the value vector size. Then, the attention between two positions i and j is computed as follows:

$$\alpha_{ij} = \text{softmax}(\frac{\mathbf{q}_i \mathbf{k}_j^\intercal}{\sqrt{d_k}}) \tag{2.4}$$

Finally, the output representation for the position $\mathbf{x}_i$ is computed as a weighted sum over the value vectors coming from all the positions in the sentence in which the weights are the computed attention values:

$$\mathbf{y}_i = \sum_j \alpha_{ij} \mathbf{v}_j \tag{2.5}$$

Multi-Head Self-Attention

Instead of only one self-attention at each layer, the authors proposed a mechanism called multi-head self-attention that would perform many self-attentions in parallel. Each self-attention output is referred to as a *head*, which is then combined with other heads by a linear projection to construct the final output:

$$\text{MH} = \text{Concat}(head_1, ..., head_h)\mathbf{W}_o \tag{2.6}$$

where h is the number of heads and $\mathbf{W}_o \in \mathbb{R}^{hd_v \times d_m}$ is the projection matrix. The transformer implements multi-head attention in three different ways. The $\mathbf{k}$, $\mathbf{q}$, and $\mathbf{v}$ vectors are calculated from the representations of the previous layer in the encoder's self-attention layer. The masked self-attention layer in the decoder is similar to the one in the encoder, with the exception that attention is only computed over prior positions and future positions are masked. In the encoder-decoder attention, $\mathbf{q}$ is computed from the previous layer, and the $\mathbf{k}$ and $\mathbf{v}$ vectors are computed from the output representations of the encoder.

Position-Wise Feed-Forward

The second sub-layer of the Transformer layers is a fully-connected feed-forward network (FNN). As shown in Equation 2.7, this network contains two layers of linear projections, which are parameterized by $\mathbf{W}_1 \in \mathbb{R}^{d_f \times d_m}$, $\mathbf{W}_2 \in \mathbb{R}^{d_m \times d_f}$ matrices and $\mathbf{b}_1 \in \mathbb{R}^{d_f}$, $\mathbf{b}_2 \in \mathbb{R}^{d_m}$ bias vectors, and ReLU non-linear function in between.

$$\mathrm{FFN}(\mathbf{x}) = \mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2 \tag{2.7}$$

This linear transformation applies to all positions within the same layer.

Positional Embedding

The Positional Embedding (PE) was another distinctive characteristic of the Transformer model. The sequential structure of RNNs embeds information about the position of the words in the sentence indirectly. However, because all of the positions in the sequence are processed in parallel and independently of each other, the model was unable to understand information about the position of the words in the sentence. The authors established the concept of positional embedding to overcome this issue and provide information to the model about the position of the words. The positional embedding is a vector that embeds information about positions and is added to the word embedding. The elements of these positional embedding vectors for odd $(2k+1)$ and even $(2k)$ indexes in the vector are computed as follows:

$$\mathrm{PE}(pos, i) = \begin{cases} \sin(pos/10000^{2k/d_m}), & \text{if} \quad i = 2k \\ \cos(pos/10000^{2k/d_m}), & \text{if} \quad i = 2k+1 \end{cases} \tag{2.8}$$

where pos is the position of the word in the sentence, and i denotes the i-th element in the positional embedding vector.

The Transformer architecture has enabled the creation of stunning pre-trained language models. In the following, we present one of the earliest pre-trained language models that relied on this architecture and revolutionized the NLP field by achieving outstanding

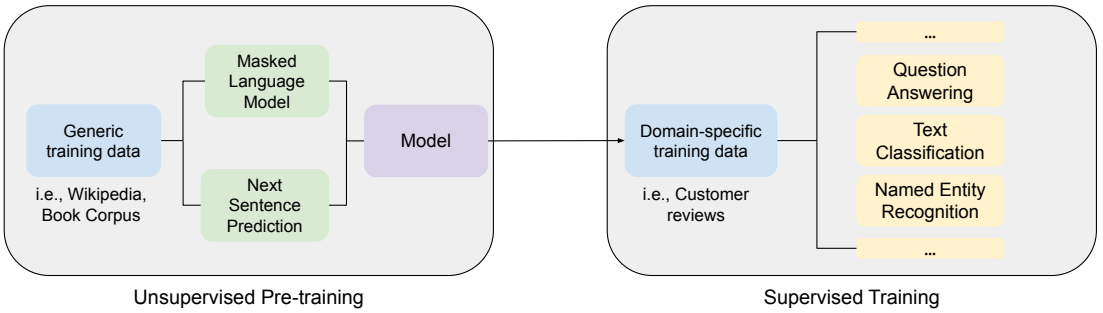

FIGURE 2.5: Pre-training and fine-tuning procedures in BERT.

results. This model has been decisive in the computational development of the tasks addressed in this doctoral thesis.

BERT

The Bidirectional Encoder Representations from Transformers (BERT) was proposed by Devlin et al. [39] in 2019. BERT along with different pre-trained language models developed since 2019 constitute the state-of-the-art methods in NLP tasks. Specifically, BERT is a transformer encoder stack (see Figure 2.4), in other words, it is constructed from transformer encoder blocks. In pre-trained languages models like BERT there are two fundamental steps: **pre-training** and **fine-tuning** (see Figure 2.5)

During **pre-training**, the model is trained on unlabeled data over different pre-training tasks. BERT uses two unsupervised training strategies: *Masked Language Modeling* (MLM) and *Next Sentence Prediction* (NSP) on a large corpus comprised of the Toronto Book Corpus and Wikipedia. The MLM task is implemented by masking 15% of the words randomly in every sentence and training the model to predict them. The NSP task is a classification task with two sentences input and the model is expected to recognize the original order between these two sentences, which increases the document level understanding. The outcome of this prior training procedure is a model capable of accurately modeling a language, and comprehending the various properties and linguistic rules of the language.

Fine-tuning is the supervised training that takes place after the pre-training of the model. The BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks (text classification, question answering, NER...). Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. At this point, the model has already acquired a statistical understanding of the language and already has some similarities with the downstream task dataset. This step requires little labeled data on a given task to specialize in it. Therefore, the amount of time and

resources needed to obtain good results are much smaller than in previous DL models like LSTM or CNN.

Some of the benefits that BERT and other pre-trained linguistic models bring to the NLP community include:

- The novel attention mechanism allows learning contextual relations between words or sub-words in a text.
- High model performance over previous ML methods.
- They are not sequential, unlike RNNs, therefore the training procedure can be easily parallelized, allowing for the training of larger models and the processing of larger volumes of text and language.
- Capabilities to fine-tune data to the specific language context and task.
- There is less need for annotated data due to the pre-training phase.

2.5.2.1 Transformer-based models for offensive language detection

After describing the groundbreaking Transformer architecture and one of the first models that make use of it, we are going to introduce different studies that used Transformer-based models to carry out the offensive language detection task.

The following studies do not only rely on BERT but also on other Transformer-based models that have been developed since the advent of the Transformer architecture, including Multilingual BERT (mBERT) [39], the cross-lingual XLM [73], and others. The initial works which employ these models were part of the OffensEval competition [5] which focused on detecting English offensive tweets, predicting the type of offense, and identifying the target to whom the offense is directed. Liu et al. [74] used the BERT model and ranked first in the identification of offensive tweets by obtaining an F_1-score of 0.83. As stated by the authors, labeled data is always limited and needs expensive human labor, hence transfer learning is always a good alternative to use. Another series of early works that began to adopt these new Transformer-based models were part of the HatEval competition [27] at SemEval 2019 workshop to identify tweets against women and immigrants in both English and Spanish. Gertner et al. [75] presented a method for adapting the pretrained mBERT model to Twitter data using a corpus of tweets collected during the same time of the HatEval training dataset achieving an F_1-score of 0.49 in English and 0.73 in Spanish in HS detection. Rozental and Biton [76] participated in both HatEval and OffensEval challenges by proposing an architecture

that they called "Multiple Choice CNN". This architecture used an ensemble of CNN including the BERT model for extracting the contextualized embeddings. They ranked 4th in the HatEval competition (Spanish task) by obtaining an F_1-score of 0.54 and 2nd in classifying the offense type in OffensEval with an F_1-score of 0.79. Benballa et al. [77] explored how transformer-based models can be combined with classical handcrafted features by proposing an approach based on a feature-level Meta-Embedding to let the model choose which features to keep and how to use them. They translated the Spanish dataset to English to use the same type of features for both languages. The model proposed by the authors achieved the best results in the development phase, but not in the testing phase, where BERT performed best. Specifically, for HS detection, they obtained a 0.77 in Spanish and a 0.52 in English in terms of F_1-score on the test set of HatEval. Besides the competitions, other works also employed transformer-based models to solve the task. The following incorporates cross-lingual models to observe their performance across languages. For instance, Sohn and Lee [78] tested a multi-channel BERT model including mBERT that joins three different BERT models (mBERT, Base BERT, and Chinese BERT). They appended an adding layer after all single fine-tuning models to make a joint representation of the three BERT models obtaining an F_1-score of 0.77 in HS detection. Ranasinghe and Zampieri [79] used cross-lingual contextual word embeddings in offensive language identification projecting predictions from English to other languages like Bengali and Spanish. As a model, they used the cross-lingual transformer model XLM [73] which has been trained on 104 languages. Their results show that XLM with transfer learning outperforms all of the other methods they tested with an accuracy of 0.85 in English, including BERT and previous state-of-the-art studies that used the same datasets. Finally, Sarkar et al. [80] built the recent fBERT model which is the BERT model retrained on the largest English offensive language identification corpus which is SOLID. Authors evaluated fBERT's performance in identifying offensive content on several English datasets. Their results show that fBERT outperforms the BERT and other offensive pre-trained language models by obtaining a test set macro-F_1 score of 0.59 on the task of HS detection in HatEval, and a 0.81 on the task of offensive language detection in OffensEval.

2.5.2.2 Integrating external knowledge

Some classification tasks benefit from the incorporation of external knowledge to more accurately predict the specific task. This is particularly applicable when dealing with highly subjective tasks such as the detection of offensive language because the expression of this issue could involve the presence of different linguistic phenomena such as emotions, sentiments, sarcasm, irony, mockery, etc. In this section, we are going to describe some

of the studies that have incorporated this type of knowledge in ML systems to tackle this task.

Sentiment and emotion analysis offers a valuable tool that helps to enhance the performance of machine learning classification systems, as shown in [81, 82]. We found that sentiments and emotions are among the most common phenomena among works that incorporate external knowledge for the detection of offensive language. Recent studies have investigated the benefit of using sentiment and emotion features for this task. Martins et al. [83] used an emotional approach that combines a lexicon-based method and a machine learning system showing that the emotional knowledge contained in the text helps to enhance the accuracy of HS detection. Rodríguez et al. [84] proposed a framework to identify Facebook pages that potentially promote HS. In order to obtain the most negative posts and comments, they applied polarity and emotion analysis, based on the idea that hateful texts contain negative emotions and sentiments. Safi Samghabadi et al. [85] introduced the gated emotion-aware attention mechanism that dynamically learns the contribution of emotional knowledge and textual information to weigh the words inside a sequence. This module is incorporated into a hybrid bidirectional LSTM and CNN architecture. They showed that this approach significantly outperforms the regular attention mechanism and in particular emotional knowledge help in short and noisy textual data. Elmadany et al. [86] developed a method for automatic data augmentation and show the utility of fine-tuning pre-existing affective bidirectional Transformer models on the downstream tasks of offensive and HS. These studies support the hypothesis that affective knowledge involved in text plays an important role in the identification of offensive language, and can be used as a valuable tool for detecting such problematic content on the Web. This affective content also has been incorporated into ML systems following an MTL methodology. For instance, Farha and Magdy [87] tested an MTL system exploring the effect of adding polarity information to perform the task of offensive language identification in Arabic tweets. They based their research on the fact that HS and offensive content always bear negative polarity. Their results showed that polarity information is correlated with HS and offensive language identification. Finally, Rajamanickam et al. [88] were the first to take into account emotional features in order to gain auxiliary knowledge through an MTL framework to detect abuse in English tweets. They proposed different MTL models, and the best result was achieved by a Gated Double Encoder model based on BiLSTM encoders. Their experiments showed that emotion detection is beneficial to abuse detection tasks in the Twitter domain.

Related to Spanish, we found very few studies that incorporate this type of knowledge to predict offensive language. Frenda et al. [89] tackled the tasks of misogyny identification by presenting an approach based on aesthetic features captured by character n-grams, sentiment information, and a set of lexicons built by analyzing misogynistic tweets.

This set of features was included in an SVM algorithm and an ensemble technique, achieving promising results in comparison with the baseline SVM without including any feature. The authors mentioned that one of the main challenges in their approach is the use of linguistic devices like irony and sarcasm in misogynistic tweets. Graff et al. [90] proposed two systems, µTC, and EvoMSA to address the challenge of detecting aggressiveness in Mexican Spanish tweets. The first is a minimalistic text categorization system that can handle general text classification tasks regardless of domain or language, and the second is a two-level Sentiment Analysis architecture that uses information from different models on the current text analysis to get a final prediction by a consensus view. They placed first in the MEX-A3T shared campaign, aggression detection task [91], demonstrating the success of their methodology. Benito et al. [92] proposed a system based on linguistic features, semantic similarity with a domain-oriented lexicon, sentiments (using the sentiment vocabulary weighted by the TF-IDF measure), word embeddings, topic modeling (both LDA and hashtags) and TF-IDF n-grams of words and characters. These features were filtered and the 3000 best were selected. The ML algorithm selected for classification was linear SVM. In contrast to previous work, the authors claimed that semantic similarity and word embeddings representations did not achieve such high-performance results when compared to other domains such as sentiment analysis tasks, and they suggested that HS detection is an open challenge that requires more research into the specific characteristics of this task. Finally, Aroyehun and Gelbukh [93] used the multilingual model XLM-RoBERTa pre-trained on Twitter texts and sentiment analysis data. They showed that sentiment analysis and the social domain adaption are beneficial for the problem of offensive language detection.

Although the mentioned studies have employed sentiment and emotion analysis to contribute to the task of detecting offensive language, we note that they have not explored in depth the benefit of this knowledge in the methodologies they employ and therefore, we consider important in this thesis to analyze how this knowledge help in the detection of offensive language in addition to the exploration of other phenomena.

2.6 Research challenges and opportunities

In this section, we will discuss some of the challenges given by the offensive language detection task, as well as the limitations identified in previous works.

As we have observed throughout this chapter, the detection of offensive language is considered a complex task in the NLP area. Great efforts have been made so far to tackle this problem. However, still has a long way to go because offensive language detection is a relatively new task with a high level of subjectivity.

The proliferation of offensive language has become a worldwide concern in recent years, owing to the vast volume of uncontrollable data being shared on social media today. However, most of the research to solve this problem has been focused on English, leaving other languages in second place. As a result, whereas English has a great number of language resources and pre-trained NLP systems, other languages have a substantial shortage of such resources. Some studies have attempted to solve this problem by simply translating texts from English or adapting developed systems from English to other languages. Unlike other tasks in NLP, offensive language may have strong cultural, demographic, and social implications which we believe should be considered for a specific language. For instance, Spanish is a rich language that presents diverse characteristics such as the frequent use of polysemy (the coexistence of many possible meanings for a word or phrase.). The Real Academia Española (RAE) dictionary contemplates several meanings for the word "zorra". The first alludes to the female fox; the second, to a "low and strong cart for transporting heavy weights"; and the third is "prostitute" which is commonly used in an offensive context. Another peculiarity of Spanish is that the vocabulary varies across the regions and even more so if we refer to the variants of Spanish in South America. For instance, although in Spain the verb "coger" means, according to the RAE, "to grasp, grab or take something or someone", in many South American countries, such as Argentina, Bolivia, Costa Rica or Nicaragua, it is a synonym for sexual intercourse. Therefore, it remains to be seen how far established approaches to offensive language detection examined in English are equally effective in other languages such as Spanish.

Regarding the social media platforms where offensive language spreads easily, we notice that Twitter is one of the platforms where the majority of research to tackle this issue is undertaken. However, social networks are comprised of several platforms, and this behavior is disseminated throughout them at the same time. As a result, throughout this PhD thesis, we attempt to investigate this phenomenon not only on Twitter, but also on other platforms such as YouTube, Instagram, and even comments posted in newspapers. It will allow us to observe how the expression of offensive language changes across these platforms since, for instance, comments posted in newspapers tend to be more formal than those written on social network platforms.

An important aspect to consider when developing computational systems for the detection of offensive language is the study of the linguistic phenomena that take place in its expression. So far, most studies have either addressed offensive language detection as a single optimization task or have incorporated affective knowledge from sentiments. However, we find that the research that has used this information to detect offensive language has not explored in detail how this knowledge benefits the task. Another significant shortcoming noticed is that, while sentiment has been extensively exploited to

detect offensive language, other phenomena involved in the expression of this problem have received little or no attention. Wiegand et al. [11] pointed out that the expression of offensiveness implies both explicit and implicit phenomenon (i.e., offensive language that is not conveyed by "unambiguously" abusive words like *dumbass*, *bimbo*, *scum*) and, in particular, they focused on identifying different subtypes of implicit abusive in existing datasets and previous work. They proposed a typology of implicit abuse[7] that includes the concepts of *stereotypes*, *perpetrators*, *comparisons*, *dehumanization*, *euphemistic constructions*, *call for action*, *jokes*, *sarcasm* and *rhetorical questions*, among others. Therefore, this study opens new directions with respect to how to approach the detection of offensive language and it has inspired us to propose the main methodology conducted in this doctoral thesis which relies on integrating different linguistic phenomena in a comprehensive computational system for detecting offensive language more accurately. This methodology will be described in detail in Chapter 5: *"Combining linguistic phenomena through a multi-task approach"*.

[7]We assume the term abuse is a synonym of offensiveness.

Chapter 3

Preliminary research on offensive language detection

This chapter constitutes the preliminary research conducted in this doctoral thesis. Specifically, the two preliminary works that have played a decisive role in the development of the doctoral thesis are described.

3.1 Introduction

The research presented in this chapter aims to understand the capabilities of both traditional NLP methods and Tranformer's emerging language models for HS detection.

In the first work, we apply for the first time traditional NLP techniques which attempt to detect misogyny and xenophobia in social media texts including supervised (traditional ML algorithms and DL models) and unsupervised learning (lexicon-based method), along with the generation of two basic lexical resources for Spanish. To the best of our knowledge, this work is one of the first in the NLP community to address the identification of both misogyny and xenophobia behaviors in Spanish.

The second work is the first attempt in this thesis to apply state-of-the-art NLP algorithms based on the Transformer architecture for HS detection in Spanish. In particular, to validate the success of these new models, we compare their performance with traditional ML models. Moreover, as we are interested to observe the importance of developing language-focused resources for offensive language detection, in this second study, we conduct experiments to compare monolingual and multilingual Transformer models for Spanish texts.

These preliminary works achieved very encouraging results, becoming the state of the art for offensive language detection in Spanish at the time they were carried out. Similarly, they have been decisive in identifying the strengths and weaknesses of current NLP models to address different tasks that involved offensive language detection such as HS, misogyny, and xenophobia detection. On the one hand, this analysis helped us to identify the scarcity of developing linguistic models in Spanish and therefore, the need to generate this type of resources essential to combat this phenomenon through NLP approaches, which constitute a fundamental part of this doctoral thesis: the generation of linguistic resources for offensive language detection in Spanish (Chapter 4: *"Resource generation"*). On the other hand, the identification of the success of Transformer language models, as well as the drawbacks observed in the NLP approaches, has been determinant to define the main NLP solution for offensive language detection proposed in this doctoral thesis (Chapter 5: *"Combining linguistic phenomena through a multi-task approach"*).

Finally, these two initial studies have been published in relevant journals as scientific papers in the NLP community [94, 95] and will be described in depth in the following sections.

3.2 Traditional methods for misogyny and xenophobia detection

In this initial work, we investigated for the first time in this doctoral thesis the performance of traditional ML and DL techniques for offensive language detection in Spanish. Specifically, we focused on the automatic detection of misogyny and xenophobia, two behaviors that have an impact on how society advances today. In addition to studying the performance of these NLP models, we identified their advantages and disadvantages, as well as the difficulties they present for the Spanish language. Moreover, in this work, we conducted the first attempt to develop a lexical resource for misogyny and xenophobia identification in Spanish.

Currently, immigrants and women are two of the most affected groups online [96]. When the HS is gender-oriented and targets women, it is referred to as misogyny, and when it is aimed at immigrants, it is referred to be xenophobia. On the one hand, social media is the primary medium for online harassment on the basis of gender [97]. This type of harassment has an impact on women's personal and professional life [98]. In fact, according to Beckman et al. [99], girls are more likely than boys to be victims of cyberbullying. In addition, different studies on sexual harassment in online video games reported that gender-based and sexual harassment are frequent in these mainly anonymous social media contexts [100, 101]. On the other hand, xenophobic HS occurs on a global scale. In

Class	Training	Development	Test
0	2,643	278	940
1	1,857	222	660
Total	4,500	500	1,600

TABLE 3.1: Number of tweets in the Spanish HatEval subsets. Class 0: non-HS, Class 1: HS.

2019, The European Commission launched a campaign with the slogan "Silence hate - Changing words changes the world" and the hashtag #silencehate to combat and prevent online HS against migrants and refugees, as well as to draw attention to the need to prevent the spread of hatred on the Internet and promote better Web use. Negative attitudes to immigration have grown in recent years, along with prejudice and more or less direct feelings of hostility towards foreigners. Anti-immigration attitudes frequently foster the spread of HS through the range of media exploited nowadays [102]. Due to the obvious massive scale data present on these platforms, automated systems based on NLP are critical for recognizing, and analyzing this type of behavior.

3.2.1 Experiments

To tackle the misogyny and xenophobia detection task, we conducted experiments based on different techniques including traditional ML algorithms, DL models, and a lexicon-based approach. In addition, in order to carry out the lexicon-based approach, we provided two new lexical resources in Spanish for identifying HS towards women and immigrants.

Dataset. To run our experiments we used the Spanish dataset provided by the organizers in SemEval19 Task 5: HatEval [27]. It contains tweets against women and immigrants. This dataset is described in detail in Chapter 2: "*Literature review*", Section 2.4: *"Corpora for offensive language detection"*. In all the experiments, we first trained a model on the training and development subsets provided by the organizers, and then we evaluated it on the test set. Table 3.1 shows the number of Spanish tweets for each HS class used in our experiments.

Dataset preprocessing. Given the inherently unstructured nature of text data, as well as the colloquial language used on the Twitter platform, it is necessary to carefully prepare the data before introducing it to the model. For this, we applied preprocessing techniques according to the language register used in social media. After tokenization, we carried out the following steps:

- Lower-case conversion data.

- Normalize URLs, emails, users' mentions, percent, money, time, date expressions, and phone numbers.
- Unpack hashtags (e.g. *#HechosReales* (#RealFacts) becomes <hashtag> *hecho reales* (Real Facts) < /hashtag>).
- Annotate and reduce elongated words (e.g. *Madree* (Motherr) becomes <elongated> *madre* (mother)) and repeat characters (e.g. !!!! becomes <repeated> !).
- Map emoticons (e.g :-) is changed to <happy>).

3.2.1.1 Traditional approaches

We applied both traditional ML and DL algorithms for misogyny and xenophobia classification.

Regarding the traditional ML methods, we chose the following classifiers: NB, SVM, LR, DT, and an ensemble voting classifier. They are described in detail in Chapter 2 *"Literature review"*, Section 2.5.1 *"Traditional methods"*. In order to apply these algorithms, we use a free software ML library for the Python programming language: scikit-learn [103].

Feature representation. The accuracy of a learning system depends on its representation of the problem. In particular, in the case of the text classification task, it is necessary to transform the document, which is mostly a string of characters, into a suitable representation for the learning classifier. Thus, in this study, we represent each document as a vector of numerical features using Frequency Term weighting (TF) which converts the text document collection into a matrix of integers generating a sparse matrix of the counts.

In this work, apart from training the different classifiers described above, we also experimented with a method based on the ensemble voting classifier which is described below.

Ensemble voting classifier. Voting is one of the most straightforward ensemble learning techniques in which the decision process involves applying several classifiers. The Voting classifier combines machine learners by using a majority vote or predicted probabilities for the classification of samples. The predictions made by the sub-models can be assigned weights.

In addition to the traditional ML algorithms, we also experimented with traditional NNs that were part of the state-of-the-art at the time of this study.

Features lookup module. We define a feature vector space for training and evaluation that is composed of unsupervised vectors of word embeddings. There are freely available several pre-trained sets of vectors of word embeddings grounded in different approaches to representing the context of a word. We specifically used the set of pre-trained vectors of word embeddings of FastText trained on Wikipedia. These vectors in dimension 300 were obtained using the skip-gram model described in [104] with default parameters.

Model architecture. Our system is based on the use of a specific gated architecture of Recurrent NN, namely LSTM [38]. This model is described in detail in Chapter 2 *"Literature review"*, Section 2.5.1 *"Traditional methods"*. In order to avoid overfitting, we add a dropout layer after each fully connected layer with a dropout rate value of 0.5. The training of the network was performed by the minimization of the binary cross-entropy function, and the learning process was optimized with the Adam algorithm [105] with its default learning rate. The training was performed following the mini-batches approach with a batch size of 32, and the number of epochs was set to 10.

3.2.1.2 Lexicon-based approach

The final approach we used to address the task of misogyny and xenophobia detection is lexicon-based. Since there were no lexical resources for misogyny and xenophobia detection in Spanish at the time of this study, we attempted to create two basic lexical resources for Spanish, one for misogyny detection and the other for xenophobia detection. After the creation of these resources, we developed a heuristic based on a lexicon-based approach to detect this type of behavior in tweets.

Lexicon Building. The methods for generating lexicons fall into two main categories, dictionary, and corpus-based approaches. The dictionary-based method consists of taking a set of words manually with the orientation (seeds) and increasing the number of words through the use of a dictionary or knowledge base (Lexical Knowledge Base-LKB). Lexical and semantic relationships are used in the search for words with affect or polarity in the LKB. This method has its limitations in finding words with specific orientations for specific domains. The corpus-based method resolves this deficiency. While different techniques have been employed, most of them start with a list of known words and try to find other related ones in a corpus of a specific domain. Thus, words with very negative orientation in one domain (*zorra* (whore) or *cerda* (slut) used to denigrate women) might have another orientation in other domain (*zorra* (fox) or *cerda* (pig) in the context of animals). With this method, a lexicon is completed with words and n-grams that are more attuned to the domain. One of the techniques used to find words from the same

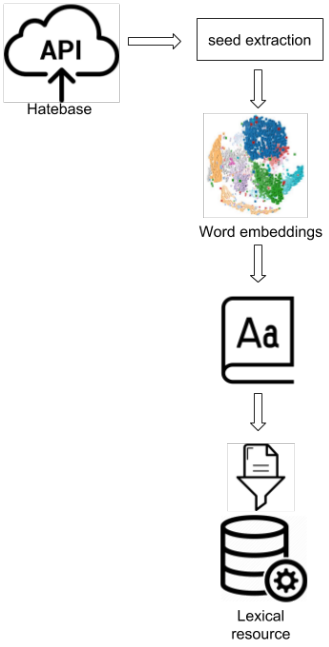

FIGURE 3.1: Scheme of lexicon building.

domain is word embedding, and it is the one we employed in our lexicon development procedure.

Building a lexical resource is an important research task in NLP in applying both supervised and unsupervised learning algorithms. We developed two linguistic resources for a lexical representation of HS knowledge about two targets (women and immigrants). Our work focused on the creation of a resource that contains a set of hateful concepts correlated with hateful words towards women and immigrants. The semantic of hate not only includes typical opinion words with negative and positive polarities but also employs rhetorical figures of speech i.e., similes and metaphors. Due to the expansion of Spanish in America and its evolution during the last five centuries, these rhetorical figures in the language are richer and more extensive, since each Spanish-speaking country has its own terms for expressing hatred. The general scheme of lexical resource building can be seen in Figure 3.1. In order to generate the best possible lexical resources, we used a hybrid approach. First, from some initial seeds, we used a specific corpus to enrich them and secondly we employ online dictionaries to complete the list of words obtained in the previous step. In addition, each step has been manually reviewed by two people.

To generate the HS lexicon towards **women** we conducted the following steps:

1. From the website `hatebase.org`, we selected five Spanish seeds for HS towards women in Spanish (*lagartona*, *perra*, *puta*, *tierrosa* and *zorra*).

2. In order to alleviate data sparseness, we used word embedding with these initial seeds. In particular, for each seed, we searched for the most similar words in the SBW word embedding model corpus and embeddings [106]. This is a pre-trained model generated using the word2vec algorithm [107] from a collection of Spanish texts with approximately 1.5 billion words. Finally, we only selected some words from the seed *puta* because usually, the other words made reference to the animal domain.

3. After that, we employed an online application[1] for extracting synonymous words and we chose those new words that had not yet been repeated. A total of fourteen words were extracted in this step.

4. The hateful words most representative of the set were searched for at `wiktionary.org` and were also included.

A total of 183 words compound the HS lexicon towards women, we refer to this lexicon as the misogyny lexicon.

On the other hand, to generate the HS lexicon towards **immigrants**, we followed the next steps:

1. From the website `hatebase.org`, we selected six seeds for HS speech towards immigrants in Spanish (*gabacho, mojado, moro, payoponi, polaco, sudaca, zambo*).

2. We used the word embeddings with some initial seeds and we searched for the most similar words in the embedding model mentioned above.

3. After that, we employed the `enciclopedia.us.es` to look for colloquial and xenophobic names.

4. The HS words most representative of the set were searched for at`wiktionary.org` and were also included.

A total of 44 words compound the HS lexicon towards immigrants.,we refer to this lexicon as the xenophobia lexicon.

In order to contribute to the problem of HS identification in Spanish towards women and immigrants, we make both lexicons publicly available in a GitHub repository[2].

Term-based Patterns. For improving our final system we analyzed some expressions including hate terms and we realized that sometimes when they are combined with other

[1] sinonimosgratis.com

[2] `https://github.com/fmplaza/hate_speech_spanish_lexicons`

Word	**expressions list *word***
puta	*puta madre* (fantastic) *puta ama* (fucking great woman) *hijo de puta* (whoreson) *hijos de puta* (whoresons) *puta boca* (fucking mouth) *puta vez* (fucking time) *puta idea* (fucking idea) *puta mierda* (fucking shit)
perra	*hijo de perra* (son of a bitch) *hijos de perra* (sons of a bitch)

TABLE 3.2: Spanish expressions with the words *puta* and *perra*.

terms the sense completely changed. For example, the misogyny lexicon includes the word *puta* (bitch). This word can be used in colloquial phrases or expressions with very positive polarity and in other cases, aimed at men with negative polarity as Table 3.2 shows. Also, there is another word *perra* (bitch) that in some expressions expresses hatred towards men. In order to avoid an erroneous classification, we have created some rules described in Algorithm 1 to be considered in our system in these special cases.

Algorithm 1: Detect misogynistic HS

Input: *d* : dataset, *ml*: misogyny lexicon, *el_puta* : expressions list puta, *el_perra* : expressions list perra
Output: HS_women

```
for each tweet in d do
    HS_women = 0;
    for each word in tweet do
        if word matches ml then
            HS_women=1;
        end
        if word = "puta" and el_puta matches tweet then
            HS_women = 0;
        end
        if word = "perra" and el_perra matches tweet then
            HS_women = 0;
        end
    end
end
```

When analyzing the HatEval corpus, we found that the use of words to communicate hatred towards immigrants is usually done in two ways: On the one hand, through

infamous and discriminatory words that identify the immigrant sector (xenophobia lexicon), and on the other hand, the use of words to indicate their nationality or words synonymous with immigrants (immigrant lexicon) followed by an insult (insult lexicon) or negative words. Therefore, we created two new different bags of words to take into account these aspects. To build the immigrant lexicon, we used the words found on the Web page `wiktionary.org`, and for the insult lexicon, we join the words found on a specific website[3] and in a GitHub repository[4]. In order to determine whether negative or positive words were being used we employed the iSOL lexicon [108] to identify them. This is a general-purpose lexicon for sentiment analysis that consists of 8,135 Spanish opinion words, 2,509 positive words, and 5,626 negative words. In Algorithm 2, we describe the rules applied in order to improve the classification of HS towards immigrants.

Algorithm 2: Detect xenophobic HS

Input: *d* : dataset, *xl*: xenophobia lexicon, *iml*: immigrant lexicon, *inl*: insults lexicon, *neg_words*: negative words iSOL, *posit_words*: positive words iSOL
Output: HS_immigrants

```
for each tweet in d do
    HS_immigrants = 0, neg = 0, pos = 0;
    for each word in tweet do
        if word matches xl then
            HS_immigrants = 1;
        else
            if word matches iml then
                if neg_words or inl matches tweet then
                    neg += 1;
                end
                if posit_words matches tweet then
                    pos += 1;
                end
            end
        end
    end
    if pos or neg != 0 then
        if neg ≥ pos then
            HS_immigrants = 1;
        else
            HS_immigrants = 0;
        end
    end
end
```

[3]https://bit.ly/3mnkjbe
[4]https://bit.ly/3NtEsrZ

Classifier	Acc	P (1)	P (0)	R (1)	R (0)	F_1 (1)	F_1 (0)	P (avg)	R (avg)	F_1 (avg)
Lexicon	0.691	0.617	0.749	0.659	0.713	0.637	0.730	0.683	0.686	0.683
LSTM	0.706	0.618	0.796	0.755	0.672	0.679	0.729	0.707	0.713	0.704
DT - unigrams	0.683	0.603	0.751	0.674	0.688	0.637	0.718	0.677	0.681	0.677
DT - bigrams	0.685	0.618	0.732	0.62	0.731	0.618	0.732	0.675	0.675	0.675
DT - uni + bi	0.694	0.624	0.746	0.648	0.727	0.636	0.736	0.686	0.688	0.686
SVM - unigrams	0.697	0.608	0.788	0.747	0.662	0.670	0.719	0.698	0.704	0.695
SVM - bigrams	0.711	0.644	0.761	0.668	0.740	0.656	0.750	0.702	0.704	0.703
SVM - uni + bi	0.719	0.632	0.806	0.764	0.688	0.692	0.742	0.719	0.726	0.717
MultinomialNB - unigrams	0.696	0.640	0.732	0.602	0.763	0.620	0.747	0.686	0.682	0.684
MultinomialNB - bigrams	0.706	0.647	0.746	0.632	0.757	0.639	0.751	0.696	0.697	0.695
MultinomialNB - uni + bi	0.734	0.664	0.79	0.718	0.745	0.69	0.767	0.727	0.731	0.728
LR - unigrams	0.711	0.628	0.789	0.736	0.694	0.678	0.738	0.709	0.715	0.708
LR - bigrams	0.736	0.693	0.763	0.647	0.8	0.669	0.781	0.728	0.723	0.725
LR - uni + bi	0.733	0.653	0.806	0.753	0.719	0.7	0.76	0.729	0.736	0.73
Vote - unigrams	0.711	0.685	0.724	0.555	0.821	0.613	0.77	0.705	0.688	0.691
Vote - bigrams	0.732	0.728	0.734	0.561	0.853	0.634	0.789	0.731	0.707	0.711
Vote - uni + bi	**0.754**	**0.721**	**0.774**	**0.658**	**0.821**	**0.688**	**0.8**	**0.747**	**0.739**	**0.742**

TABLE 3.3: Results achieved by the lexicon-based approach, the DL model and the traditional ML classifiers. P: Precision, R: Recall.

Given that the objective of the task was to identify whether a tweet contains HS towards women or immigrants, in order to perform the classification we added up the value of HS_women and HS_immigrants and finally, if the sum is not 0, we labeled it as hateful (HS = 1).

3.2.2 Results and discussion

The results of all our text classification experiments are presented in Table 3.3. They have been evaluated using the usual metrics for text classification, including accuracy, precision, recall, and F_1-score. Let us start with the model that we consider as our baseline, which is the lexicon-based approach. It should be noted that the results obtained (F_1: 0.683) are almost the same as those of the DT classifier with the combination of unigrams and bigrams (F_1: 0.686). Therefore, the linguistic resources generated and the rules applied in the lexicon-based system have achieved more than acceptable results, even compared with some traditional ML algorithms. One of the main advantages of using this method is that there is no need for labeled data as the learning procedure is not necessary. However, some of its limitations are that it requires powerful linguistic resources which are not always available, and it is also difficult to take the context into account.

Now it is time to review the results of our experiments using the traditional ML algorithms. In Table 3.3 the results obtained by each of the classifiers can be observed using different features: TF-unigram, TF-bigram, and a combination of them. We notice that the Macro F_1-score value obtained using the combination of unigrams and bigrams is

better than the one obtained using unigrams and bigrams separately. We assume that this is because, in Spanish, there are some colloquial phrases or expressions like for example those shown in Table 3.2 with very positive polarity in bigrams but the single words (unigrams) that compose them are a clear example of HS in misogyny. Therefore, the classifier benefits from the knowledge of bigrams. If we look at the results of each classifier, we see that the best performance is achieved by the LR, with a Macro-F_1 score of 0.73 with the combination of unigrams and bigrams. On the other hand, the DT classifier performance is the one that obtains the worst results among the traditional ML algorithms. The results show a very good performance of two classifiers: MultinomialNB and LR with the combination of unigrams and bigrams. For this reason, we decided to ensemble them in a voting classifier. We have established the voting parameter to "hard" which means that the model used predicted class labels for majority rule voting. Every individual classifier votes for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels. A general scheme of the system can be seen in Figure 3.2. This scheme performs robustly better in the case of the combination of unigrams and bigrams than when classifiers are used separately. This is due to the fact that when one of the two classifiers does not make the correct prediction, the final prediction is corrected by the other. Some of the advantages of using this type of approach are the following: it does not need a large training set in order to achieve good results, and in addition, the decisions made by the classifiers are easy to interpret. However, it is not flexible enough to capture more complex relationships naturally.

With respect to the results obtained with the DL approach, we observe that it does not improve the results obtained by the voting classifier. Traditional DL text classification techniques in general need a large amount of training data and that is what, at the time of this study, was not available for the Spanish HS detection task.

Table 3.4 shows the most informative unigrams and bigrams for each class (misogyny and xenophobia) in the case of the best model classifier (Vote). As can be observed, most of the unigrams are insults to women and immigrants. In the case of bigrams, we notice that verbs are often used in imperative mode (expressing an order) with an insult (*cállate zorra* (shut up bitch)) for the misogyny class. For the xenophobic class, negative adjectives are used together with a derogatory name referring to the nationality of the immigrant (*malditos sudacas* (damned South Americans), *putos moros* (fucking Moors), *putos sudacas* (fucking South Americans), *malditos árabes* (damned Arabs)). In conclusion, we would like to emphasize that bigram features are very important for HS detection in Spanish due to the reason that people who speak this language not only tend to use insults to express hate, as may be the case with English, but also often use some expressions such as those shown in Tables 3.2 and 3.4). In addition, as we

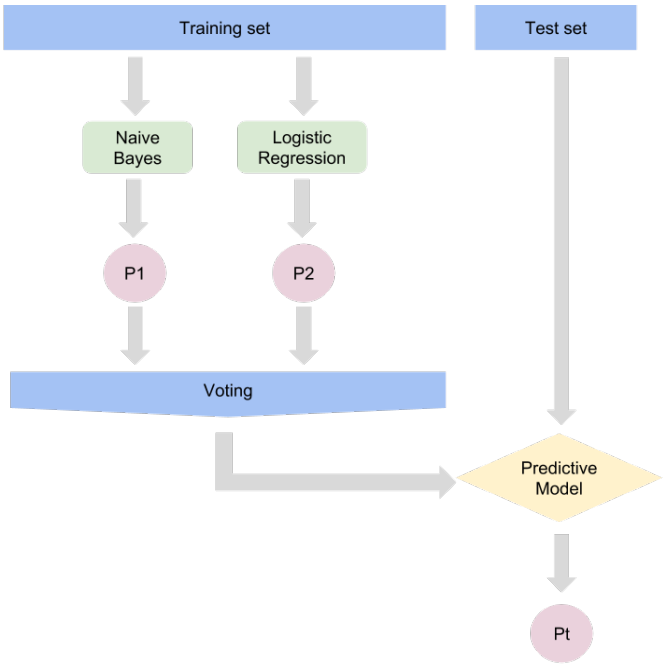

FIGURE 3.2: Scheme of the final system.

	Misogynist class	Xenophobic class
Unigrams	*fregar* (wash)	*moromierda* (piece of shit moor/arab)
	zorra (bitch)	*sudaca* (spic)
	puta (whore)	*indio* (indian)
	pendeja (slut)	*moro* (moor/arab)
	feminazi	*negrata* (nigger)
	perra (bitch)	*escoria* (scum)
	cómemela (suck my dick)	*putos* (fuckers)
	viejas (old biddies)	*flojo* (weak)
	pelotuda (stupid)	*pendejo* (berk)
	gorda (fatty)	*gabacho* (frenchy)
Bigrams	*cara zorra* (bitch-face)	*palestino indocumentado* (undocumented palestinian)
	de perra (bitch)	*negro indocumentado* (undocumented nigger)
	cállate puta (shut up, bitch)	*malditos sudacas* (damn spics)
	la puta (the bitch)	*los árabes* (arabs/moors)
	zorra amiga (bitch friend)	*eres escoria* (you are scum)
	cállate hija (shut up, daughter)	*puta rata* (rat whore)
	cállate zorra (shut up, bitch)	*putos moros* (fucking arabs/moors)
	qué guarra (what a slut)	*putos sudacas* (fucking spics)
	hostia puta (shit/fucking hell)	*los negratas* (niggers)
	puta vez (fuck's sake)	*malditos árabes* (damn arabs/moors)

TABLE 3.4: Most informative words and bigrams for each class (misogynist class, xenophobic class).

Classifier	Training (number of tweets)	F_1 (avg)
Lexicon	-	0.683
Vote - uni + bi	500	0.628
Vote - uni + bi	1000	0.670
Vote - uni + bi	1500	0.706
Vote - uni + bi	5000	0.742

TABLE 3.5: A comparative of test results between lexicon-based approach and vote with different training set sizes.

have mentioned before, bigrams can completely change the polarity of one of the words that compose them. If we compare lexicon-based and supervised learning approaches, we notice that in both cases bigram features help to improve the classification. In fact, some rules based on expressions used by the lexicon-based approach are also used by the supervised learning approach.

To highlight the need to develop lexical resources, we compared the best vote approach (uni + bi) and the lexicon-based approach by testing different sizes of the training set. Table 3.5 shows the performance of the vote system considering 500, 1000, 1,500 and 5,000 tweets for training. As we can see, if we reduce the training set, we noticed that the lexicon-based method begins to be better in the 1500 to 1000 tweets gap. We consider that this is because the system based on supervised learning does not have enough examples to train and consequently worse results are obtained. Specifically, in this type of task (HS) is not easy to gather large training sets. Therefore, the lexicon-based method could be very useful when it is not possible to have a large enough dataset to train a supervised system.

As mentioned in Chapter 2: *"Literature Review"*, Section 2.3 *"Shared task evaluation campaigns"*, SemEval organized a task on HS detection against immigrants and women named HatEval. Some representative results achieved by the participants are shown in Table 3.6. Noteworthy are the top three teams Francolq2 [109], Luiso.vega [110] and Gertner [111]. As we can see, the Macro F_1-score value of the baseline system (rank = 21) and the best system (rank = 1) is between 0.701 and 0.73 hence the improvements between the different systems are slightly noticeable. This shows that it is a complex task since achieving such a considerable improvement in the base system is not easy. It should be noted that our results obtained with the lexicon-based system are not noticeably worse than the baseline system of the task.

In conclusion, the results show that the ensemble schema is performing very well in detecting HS against women and immigrants. We would like to mention that our ensemble classifier system outperforms the baseline system and the best system of the HatEval

User name (rank)	Acc	P (1)	P (0)	R (1)	R (0)	F_1 (1)	F_1 (0)	P (avg)	R (avg)	F_1 (avg)
Our proposal	**0.754**	**0.721**	**0.774**	**0.658**	**0.821**	**0.688**	**0.8**	**0.747**	**0.739**	**0.742**
francolq2 (1)	0.731	0.639	0.829	0.8	0.683	0.711	0.749	0.734	0.741	0.73
luiso.vega (2)	0.734	0.655	0.804	0.748	0.723	0.699	0.761	0.729	0.736	0.73
gertner (3)	0.729	0.622	0.878	0.876	0.626	0.727	0.73	0.75	0.751	0.729
geoint (20)	0.702	0.606	0.816	0.795	0.636	0.688	0.715	0.711	0.716	0.701
SVC baseline (21)	0.705	0.623	0.779	0.72	0.695	0.668	0.735	0.701	0.707	0.701
Halamulki (22)	0.703	0.618	0.784	0.732	0.682	0.67	0.729	0.701	0.707	0.7
vista.ue (38)	0.612	0.532	0.66	0.486	0.7	0.508	0.679	0.596	0.593	0.594
bogdan27182n (39)	0.546	0.47	0.723	0.8	0.367	0.592	0.487	0.597	0.584	0.54
DA-LD-Hildesheim (40)	0.511	0.405	0.582	0.392	0.595	0.398	0.588	0.493	0.494	0.493

TABLE 3.6: Some systems results by the participants in Spanish HatEval task. P: Precision, R: Recall.

System	Errors	Predicted 1	Predicted 0
Lexicon-based	496	271 (54,63%)	225 (45,37%)
Voting	395	169 (42,78%)	226 (57,22%)
LSTM	453	192 (42,38%)	261 (57,62%)
All (in common)	106	36 (34%)	70 (66%)

TABLE 3.7: Number of instances mislabeled by each system, broken down by wrongly assigned label.

SemEval task by a substantial margin, which demonstrates that our best system is a successful methodology for detecting HS toward immigrants and women.

3.2.3 Error analysis

The main purpose of this section is to carry out an error analysis to identify the weaknesses of our different approaches. For each system, we checked the instances in the test set that were wrongly labeled. Moreover, we also analyzed the instances that were wrongly labeled by all three of them.

The three different approaches (lexicon-based, voting, and LSTM) predicted the same wrong labels 106 times out of 1,600. The results showing the percentages by wrongly assigned labels for each system are summarized in Table 3.7.

The common errors are highly biased towards false negatives, except in the case of the lexicon-based system.

Four examples, respectively two false positives and two false negatives, are:

- *Fútbol sudaca de mi vida.* (Football spic of my life).

- *Esta chica es puta maravilla en un escenario... sin más.* (This girl is a fucking wonder on a stage... no more).
- *Estoy escuchando una puta canción y la pelotuda de Demi Lovato se pone a hablar en el medio. CANTÁ Y CALLATE LA BOCA.* (I am listening to a fucking song and that asshole Demi Lovato starts talking in the middle of it. SING AND SHUT YOUR MOUTH).
- *200 inmigrantes saltan la valla de Ceuta y hieren a 7 agentes. No os olvidéis ahora darles techo, comida, agua y una paguita mensual.* (200 immigrants jump the fence of Ceuta and injure 7 agents. Don't forget now to give them shelter, food, water and a monthly payment). qué descanses, buenas noches sueña con los angelitos y no tengas pesadillas con tantos juais.

In the first false positive a negative word *sudaca* (spic) is used humorously, for the purpose of praising South American football. However, our system misclassifies it because the word *sudaca* (spic) is commonly used in a derogatory way to refer to a person from South America. In the second false positive, an expression *puta maravilla* (fucking wonder) is used in a positive way. However, our system misclassifies it because this expression contains the word *puta* (whore).

In the first false negative, a misogynistic message is expressed, although covertly, implying that the target should *CANTÁ Y CALLATE LA BOCA* (SING AND SHUT YOUR MOUTH). In the second false negative, the message contains irony, a linguistic phenomenon difficult to identify with systems.

As a result, we identified in our investigation that the models experienced significant difficulty when detecting offensive language. On the one hand, the models generate false positives when there are insults with positive connotations or when swear words are used in positive sentences. On the other hand, the model generates false negatives when there is a lack of context in the instance, linguistic phenomena such as irony or sarcasm are used, and offensiveness is implicitly stated.

Finally, we would like to mention that we found some incorrectly labeled messages in the dataset, which is a potential source of confusion for a classifier.

Class	Retrieved tweets	Selected tweets	Labeled tweets
0	-	-	4,433
1	-	-	1,567
Total	2 million	8,710	6,000

TABLE 3.8: Number of tweets in Spanish HaterNet dataset.

3.3 Traditional methods vs Transformers for hate speech detection

In the previous study, we could observe that traditional ML systems are good estimators of two main problems related to HS detection: misogyny and xenophobia. With this study, we are interested to investigate the new era of Transformer-based models for HS detection in Spanish and compare their performance with the traditional methods in order to provide a deeper understanding of the capabilities of new techniques based on transfer learning. Moreover, we aim to observe if offensive language detection is language-dependent by comparing the performance of monolingual models trained specifically for Spanish and multilingual models trained in different languages.

3.3.1 Experiments

Datasets. We evaluated our experiments on two available Spanish datasets composed of tweets that may contain HS. The first one, HatEval, is the one used in the first study mentioned above and the second one we incorporate in this study is HaterNet. Table 3.8 shows the number of retrieved, selected, and manually labeled tweets for the creation of HaterNet. For a detailed description of this dataset see Chapter 2: "*Literature review*", Section 2.4: *"Corpora for offensive language detection"*.

For our experiments, in the case of HatEval dataset, the union of training and development sets builds the training set which contains 2,921 not hateful tweets and 2,079 hateful tweets. The test set contained 940 not-hateful tweets and 660 hateful tweets. For HaterNet, we performed 10-fold cross-validation since this dataset is not available with partitions.

Dataset preprocessing. To perform the preprocessing procedure in both HatEval and HaterNet datasets we carried out the following steps:

- Converting all tweets to lower case.

- Normalising URLs, emails, users' mentions, percent, money, time, date expressions, and phone numbers.
- Annotating and unpacking hashtags splitting the hashtag to its constituent words (e.g., #ILoveAnimals becomes <hashtag> I Love Animals < /hashtag>).
- Annotating and reducing elongated words (e.g. hateeee becomes ¡elongated¿ hate) and repeat words or punctuations (e.g. !!!! becomes <repeated> !).

We conducted experiments based on different approaches that we establish as our baselines. First, we evaluated traditional ML and DL models including LSTM, CNN, and Bi-LSTM, SVM, and LR. We establish them as our baselines in our experiments. Then, we evaluated recent pre-trained language models based on the Transformer mechanism using multilingual models including BERT, XLM, and a monolingual one (BETO) available for Spanish. Specifically, we rely on these models because of their major advantages: they do not need a large dataset, not always available, specifically for languages other than English; they are able to capture long-term dependencies in language, and they effectively incorporate hierarchical relations which is very important in languages like Spanish due to its syntactic and semantic complexity. Finally, we analyze the comparison of the Transformer-based models' performance with our baseline systems and with the latest state-of-the-art results in HS detection for Spanish.

3.3.1.1 Traditional approaches

Regarding the traditional methods, we considered in this study both ML and DL traditional classifiers to carry out the task of HS detection. For ML, we chose the two models most commonly used for classification tasks at the time of this study: SVM and LR classifiers. For a detailed description about these models, please see Chapter 2 *"Literature review"*, Section 2.5.1 *"Traditional methods"*. In order to extract features for their inclusion in each traditional classifier, we use two types of text representation: the TF-IDF scheme and word embeddings. We specifically choose the set of 300-dimensional pre-trained vectors of word embeddings of fastText trained on a Spanish Unannotated Corpora[5].

Regarding the NN approach, we used two traditional DL classifiers, the CNNs, and the Recurrent Neural Networks. These networks are explained in detailed in Chapter 2 *"Literature review"*, Section 2.5.1 *"Traditional methods"*. we use the same embedding model that we use for traditional ML algorithms and is described above.

[5] `https://bit.ly/3mB592c`

Dataset	Hyperparameter	Options	LSTM	BiLSTM	CNN
HaterNet	Size	[50, 100, 150]	150	50	150
	Dropout	[0.25, 0.5]	0.25	0.25	0.5
	Activation	[tanh, relu]	relu	relu	relu
	Optimizer	[Adam, SGD]	Adam	Adam	Adam
	Batch size	[8, 16, 32, 64, 128, 256]	32	64	32
	Learning rate	[0.001, 0.002, 0.01, 0.02]	0.01	0.01	0.002
HatEval	Size	[50, 100, 150]	150	150	100
	Dropout	[0.25, 0.5]	0.25	0.25	0.25
	Activation	[tanh, relu]	relu	tanh	tanh
	Optimizer	[Adam, SGD]	Adam	Adam	Adam
	Batch size	[8, 16, 32, 64, 128, 256]	16	8	16
	Learning rate	[0.001, 0.002, 0.01, 0.02]	0.002	0.002	0.001

TABLE 3.9: Best hyperparameter values selection of the DL models.

Hyperparameter optimization. The NNs used in this study contain a number of hyperparameters that must be estimated in order to achieve optimal results. For this purpose, we use the validation set performance to select the best set of hyperparameters for the test set in the case of HatEval dataset. For HaterNet, we have split the dataset into train and test. Then, for hyperparameter tuning, we perform a 10-fold cross-validation with the training dataset, in this way, we get the best hyperparameters over 20 different combinations with Bayesian Optimization. Finally, we used the test set to predict and evaluate the predictions using the best hyperparameters. Table 3.9 shows the hyperparameter options that have been tested on each dataset and the resulting best parameters for each model (LSTM, CNN, and BiLSTM). In addition, we use early stopping as a form of regularization to avoid overfitting during supervised training of a NN, by stopping training before the weights have converged.

3.3.1.2 Transformer-based models

We experimented for the first time in this doctoral thesis with pre-trained language models based on Transformer for HS detection. Specifically, we rely on BERT [39] and XLM [73] models which were the ones available at the time of this study. BERT was originally pre-trained on English texts and then extended to other languages in the form of Multilingual BERT (mBERT). mBERT is a single language model pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages, including Spanish. There are two mBERT models available: BERT-Base Multilingual Cased and BERT-Base Multilingual Uncased. In particular, for this study, we chose the BERT-Base, Multilingual Cased checkpoint[6]. It is worth mentioning that although we use this checkpoint, we decided to convert the text to lowercase while preprocessing the dataset

[6] `https://github.com/huggingface/transformers`

since the BERT tokenization of the uppercase words for Spanish does not work properly. For example, three tokens are obtained by applying the tokenization in the word *PUTA* (WHORE) = 'P', '##UT', '##A'. This can be a potential source of confusion for the classifier since the same word can be represented in different ways. However, when the word *puta* (whore) is lowercase we get the right token *'puta'*. Therefore, it is important to normalize all words to lowercase in order to achieve a good interpretation of the words by the classifier. A drawback in the mBERT model is that it was pre-trained on the concatenation of monolingual corpora from different languages and it does not provide a language detection mechanism, so the word piece tokenizer can occasionally confuse languages. Moreover, it does not have any explicit procedure to encourage translation equivalent pairs to have similar representations. For this reason, we opted to run our experiments with a recent monolingual BERT model called BETO[7][112] trained specifically on Spanish data. The comparison of multilingual and monolingual models will allow us to determine whether a model trained solely for Spanish performs better than a multilingual model. Another multilingual model that we employ in our study is XLM which has achieved ground-breaking success in cross-lingual classification, unsupervised machine translation and supervised machine translation tasks. BERT has not been optimized for multi-lingual models since most vocabulary is not shared between languages. In order to address this problem, XLM processes all languages with the same shared vocabulary created through a preprocessing technique named Byte Pair Encoding [113] and employs a dual-language training mechanism with BERT in order to learn the relationships between words in different languages. It uses a hidden size of 1280, 16 transformer blocks, and 16 self-attention heads. Specifically, we chose the xlm-mlm-100-1280 checkpoint[8] which covers 100 languages, including Spanish.

Hyperparameter optimization. In our experiments, we fine-tuned these Transformer-based models on the HatEval and HaterNet datasets. By fine-tuning, the model updates the weights using the annotated dataset that is new to an already trained model. While fine-tuning the model, it is recommended to experiment with the following hyperparameters: batch size, learning rate, max sequence, and number of epochs [114, 115]. Therefore, we performed hyperparameter optimization by fine-tuning with different values as shown in Table 3.10.

3.3.2 Results and discussion

In this section, we explore the capabilities and limits of the different ML approaches we have evaluated. In order to do so, we have employed the usual metrics in NLP tasks,

[7] `https://github.com/dccuchile/beto`
[8] `https://huggingface.co/transformers/multilingual.html`

Dataset	Hyperparameter	Options	mBERT	XLM	BETO
HaterNet	Epoch	[2, 3, 4]	2	3	2
	Batch size	[16, 32]	16	16	16
	Learning rate	[2e-5, 3e-5]	2e-5	2e-5	2e-5
HatEval	Epoch	[2, 3, 4]	3	4	3
	Batch size	[8, 16, 32]	16	32	16
	Learning rate	[2e-5, 3e-5]	3e-5	2e-5	2e-5

TABLE 3.10: Best hyperparameter values selection on the Transformer language models.

		Non-HS			HS			Macro-Avg		
Dataset	Model	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)	P (%)	R (%)	F_1 (%)
HaterNet	LR (TF-IDF)	79.79	**96.58**	87.31	**77.53**	30.66	43.16	78.66	63.62	65.24
	SVM (TF-IDF)	83.26	91.06	86.93	66.10	48.28	55.33	74.68	69.68	71.13
	LR (Embeddings)	80.24	94.21	86.61	68.42	34.25	44.98	74.33	64.23	65.80
	SVM (Embeddings)	80.90	92.88	86.41	66.39	38.11	47.81	73.65	65.50	67.11
	CNN	80.76	93.79	86.79	67.65	36.74	47.62	74.20	65.27	67.20
	LSTM	81.58	94.47	87.55	71.68	39.62	51.03	76.63	67.04	69.29
	BiLSTM	80.36	96.50	87.69	77.04	33.23	46.43	78.70	64.86	67.06
	XLM	85.06	89.95	87.38	66.63	55.29	59.97	75.84	72.62	73.68
	mBERT	85.03	88.68	86.76	64.65	55.80	59.33	74.84	72.24	73.05
	BETO	**87.29**	90.19	**88.66**	70.45	**62.82**	**65.80**	**78.87**	**76.51**	**77.23**
HatEval	LR (TF-IDF)	69.79	77.91	77.63	71.82	62.53	66.85	70.80	70.22	70.24
	SVM (TF-IDF)	68.83	78.52	73.36	73.18	62.24	67.27	71.01	70.38	70.31
	LR	96.50	78.80	86.76	26.52	72.81	38.88	61.51	75.80	62.82
	SVM	**97.29**	78.65	**86.98**	25.24	76.70	37.98	61.27	77.87	62.48
	CNN	76.73	**83.83**	80.12	**73.47**	63.79	68.29	75.10	73.81	74.21
	LSTM	84.87	70.43	76.98	66.10	82.12	73.24	75.48	76.27	75.11
	BiLSTM	82.21	75.21	78.56	68.51	76.82	72.43	75.36	76.02	75.49
	XLM	86.68	72.66	79.05	68.35	**84.09**	75.41	77.51	78.38	77.23
	mBERT	83.48	72.55	77.63	67.05	79.55	72.77	75.26	76.05	75.20
	BETO	86.16	74.15	79.70	69.28	83.03	**75.53**	**77.72**	**78.59**	**77.62**

TABLE 3.11: Results on the Spanish HS datasets. Best results are shown in bold. P: Precision, R: Recall.

including precision, recall, F_1-score, and the macro-average.

We compared the performance of the different models on HS in two Spanish datasets. Table 3.11 shows the prediction performances we have achieved for each classifier and each dataset. In all the models, we used word embeddings as the input features, while in the case of traditional ML algorithms, we also tested the statistical-feature TF-IDF including LR (TF-IDF) and SVM (TF-IDF). In most ML algorithms, TF-IDF produces better results than the word embeddings features, especially in the case of the HatEval dataset.

The baseline experiments (traditional ML and DL models) performed well despite the lack of sufficient training instances. In terms of macro-F_1 score, the NNs achieve better results than traditional algorithms in the HatEval dataset, but this is not the case for

HaterNet, where the best baseline is the SVM (TF-IDF) classifier. These findings are in line with the work of Zhang et al. [116] where traditional methods were found to have comparable performance to deep NNs on different sentence classification tasks.

As shown in Table 3.11, the Transformer language models (BERT, XLM, and BETO) substantially outperform the baselines systems in both datasets in terms of Macro-F_1 score. The best performance was achieved by BETO followed by the pre-trained multilingual models, XLM and mBERT which behaved in the same way in both datasets. It is important to note that for both datasets, XLM achieves better results than mBERT. One reason could be that mBERT was pre-trained on the concatenation of monolingual corpora from different languages and the tokenizer confuses them, therefore the coverage of the vocabulary found in the datasets is not very accurate.

Probably, due to the fewer number of instances in the HS class, the F_1-score gets lower results for this class in all the classification models. Specifically, for HaterNet there is a great difference in the precision and recall scores between both classes. Despite that, BETO achieves the best macro-recall and macro-F_1 scores. In the case of HatEval, although there is less difference between the results obtained in both classes, BETO also achieves the best results in terms of the macro-scores.

Table 3.12 shows, at the time of this study, the state-of-the-art results of HS detection in Spanish obtained by previous studies using HaterNet and HatEval datasets. For HaterNet dataset, Pereira-Kohatsu et al. [34] tested up to 19 different models taking into account different combinations of features, classification models, and thresholds. Their most successful SVM classifier was implemented using frequency-based features computed for unigrams, POS tags, emojis, suffixes, and expression tokens (Model 9). Our SVM classifier is implemented using the TF-IDF scheme, obtaining a 14.6% improvement in terms of F_1 for HS class. Their best model employs words, emojis, token embeddings, and TF-IDF as input features to LSTM+MLP (Model 7). Our BETO model improves on the best methodology presented by Pereira-Kohatsu et al. [34] by 7.7% in terms of F_1-score for HS class. Regarding the HatEval dataset, Sohn and Lee [78] outperformed the best result obtained in the HatEval competition [117] by testing a multi-channel BERT model including mBERT that joins three different BERT models (mBERT, Base BERT, and Chinese BERT). They appended an adding layer after all single fine-tuning models to make a joint representation of the three BERT models. Using the monolingual model BETO we surpassed the results of Sohn and Lee [78] with an improvement of 1.18%, achieving a Macro-F_1 score of 77.18%. It is worth mentioning that we also surpass the results obtained in our previous study in the frame of the preliminary research on offensive language detection of this doctoral section, described in Section 3.2: *"Traditional methods for misogyny and xenophobia detection"*. Specifically,

Dataset	System	F_1 (Non-HS)	F_1 (HS)	Macro-F_1
HaterNet	SVM (Pereira-Kohatsu et al. [34])	-	48.3	-
	LSTM+MLP (Pereira-Kohatsu et al. [34])	-	61.1	-
	BETO (Our proposal)	**88.7**	**65.8**	**77.2**
HatEval	multi-channel BERT (Sohn and Lee [78])	-	-	76.6
	Ensemble voting classifier (Plaza-del-Arco et al. [94])	**80.0**	68.8	74.2
	BERT (Gertner et al. [75])	73.0	72.7	72.9
	SVM (Vega et al. [117])	76.1	69.9	73.0
	BiGRU (Paetzold et al. [69])	77.1	52.1	64.6
	BETO (Our proposal)	79.7	**75.5**	**77.6**

TABLE 3.12: State-of-the-art results for HS detection in Spanish. Best results are shown in bold.

Dataset	System	Errors	Predicted 1 (FP)	Predicted 0 (FN)
HaterNet	mBERT	125	46 (36.80%)	79 (63.20%)
	XLM	119	34 (28.57%)	85 (71.43%)
	BETO	106	41 (38.68%)	65 (61.32%)
	All (in common)	54	14 (25.93%)	40 (74.07%)
HatEval	mBERT	432	288 (66.66%)	144 (33.33%)
	XLM	376	252 (67.02%)	124 (32.98%)
	BETO	359	255 (71.03%)	104 (28.97%)
	All (in common)	207	148 (71.50%)	59 (28.50%)

TABLE 3.13: Number of instances mislabeled by each Transformer language model.

BETO surpasses the best model (the voting classifier) by 3.4 points in terms of macro-F_1 score and 6.7 points in the HS class. This comparison demonstrates the success of the Transformer language models for Spanish offensive language detection. Finally, we highlight the importance of training a model on a Spanish corpus, since the monolingual model BETO is the model that achieves the best results in both datasets, we also highlight the importance of hyperparameter tuning to find the best combination of model hyperparameters in each dataset.

3.3.3 Error analysis

We conducted an error analysis in order to identify the weaknesses of the systems we have employed to detect HS in Spanish tweets. This could be very helpful in determining challenges in this task and provide insights into the classifier's performance.

In particular, we carried out an error analysis on the three Transformer-based models that achieved the best results for both datasets. For each model and dataset, we checked the instances in the test set that were wrongly labeled. In addition, we analyzed the instances that were wrongly labeled by all three of them.

		HaterNet				HatEval			
Model	Vocabulary	Included	%	Not Included	%	Included	%	Not Included	%
mBERT	119,547	3,635	24.98	10,914	75.02	4,025	22.81	13,620	77.19
XLM	170,871	3,820	30.67	8,637	69.33	4,221	23.92	13,424	76.08
BETO	31,002	5,912	35.52	10,729	64.48	6,974	39.51	10,671	60.49

TABLE 3.14: Vocabulary coverage by the Transformer language models

The results showing the percentages by wrongly assigned labels for each system and dataset are summarized in Table 3.13. In the case of the HaterNet dataset, the three best accurate models (mBERT, XLM, and BETO) predicted the same wrong labels 54 times out of 600 and 207 times out of 1,600 in the case of the HatEval dataset.

Regarding the best system BETO, the instances mislabeled are biased towards FP in the case of the HatEval dataset (71.03%) and towards FN in the HaterNet dataset (61.32%).

Table 3.14 shows the vocabulary coverage by the Transformer language models in the datasets evaluated. For each dataset four columns are shown: the dataset words included in the model (Included column in Table 3.14) and their equivalent percentage, the dataset words not included in the model (Not Included column in Table 3.14) and their associated percentage. It is important to remark that BETO is the model that provides the greatest coverage with 6,904 included words (40.42%) in HaterNet and 8,246 included words (37.53%) in HatEval. Specifically, we find some words from the HS datasets that are included in BETO vocabulary but not in the multilingual models' vocabulary such as *gorda* (fat), *maricón* (fag), *lesbiana* (lesbian), *perra or zorra* (bitch), *inmigrantes* (immigrants), *anormal* (abnormal), *cabrón* (bastfard), *gilipollas* (asshole) and *miserable* (miserable). All of these words have an important meaning in the context of detecting HS, including a description of the physical attributes, sexual orientation, words that mention discriminated groups, and swearwords, among others. The greatest coverage in BETO vocabulary could be one of the main reasons why this model achieves the best performance in both HS datasets. From this, we could conclude that it is important to pre-trained a model on Spanish data because although mBERT and XLM provide a larger vocabulary, they were pre-trained on the concatenation of monolingual corpora from different languages, and as we can see in Table 3.14 the coverage vocabulary is less than for BETO.

Finally, Table 3.15 shows some instances that were wrongly predicted by the best system, BETO. In particular, two FP instances and two FN instances can be seen.

In the first FP, the user is expressing disapproval of people who act in a misogynistic manner because of the way they dress. However, the BETO model misclassifies this, presumably due to the existence of the derogatory word "bitch". In the second FP, different insults towards immigrants are used (*negrata* (nigger), *maricón* (faggot), sudaca

Tweet	True label	Predicted
¿Me podéis dar más asco los que pensáis que por enseñar alguna parte del cuerpo ya te tienen que tratar como a una puta? (Can I be more disgusted by those who think that because they show some part of their body they already have to treat you like a whore?)	0	1
Claro, pero no es lo mismo que nigger o maricón, sudaca sí es, por ejemplo. (Right, but it's not the same as nigger or faggot, spic is, for example)	0	1
400 voltios y que quiten las concertinas, y el que tenga huevos que salte (400 volts and take off the concertinas, and whoever has the balls for it should jump)	1	0
Premio al insulto del año para estos muerdealmohadas! (Insult of the year award for these pillow biters!)	1	0

TABLE 3.15: Tweets mislabeled by the BETO model with the corresponding translation in English.

(spic)), but the classifier is not able to identify the context of the tweet. In this case, it is also difficult for humans to know whether the user is attacking immigrants or not because of the lack of context.

In the first FN the user is conveying a xenophobic comment, however, our model misclassifies it, most likely because there is no clear mention of immigrants in the tweet. In the second FN an insult to homosexuals is used, *muerdealmohadas* (pillow biters). However, our system is unable to label the tweet as HS, maybe because this word is not in BETO's vocabulary and there are insufficient occurrences of this word being used in the training set for the system to learn it as an insult to this protected category.

Therefore, with the error analysis of this study, we can observe that some of the most challenging cases for the classifier while detecting HS instances are: the lack of context, the presence of offensive words not learned during the training procedure, and the identification of the reference to other people's utterances.

Finally, we would like to mention that we found some incorrectly labeled messages in the dataset. In particular, we have found some tweets labeled as HS when they are offensive but not towards a protected group which is a potential source of confusion for a classifier.

3.4 Conclusion

In this chapter, we presented two pivotal works completed at the beginning of this doctoral thesis. The first study is one of the pioneering works in the NLP field focusing

on detecting HS in Spanish with traditional methods. Three different approaches are explored: a lexicon-based approach, a supervised ML approach, and a DL approach. For the lexicon-based approach, due to the lack of resources concerning women and immigrants in Spanish, we conducted the first attempt to create linguistic resources for Spanish HS detection in order to apply some patterns to the classification of the dataset. The results achieved are comparable to the baseline system for the SemEval 2019 HatEval task and to the traditional ML classifier, DT. This is a remarkable achievement, as we make two basic lexical resources available to the NLP community and demonstrate that lexical-based methods can be useful for the detection of this phenomenon in Spanish. For the traditional ML approach, we have explored several classifiers, as well as the use of n-gram features. It has been observed that when we apply as a feature the combination of unigrams and bigrams the Macro F_1-score increases in all classifiers, mainly due to the use of expressions while expressing HS in Spanish. This voting schema of the two best classifiers outperforms not only all the other systems described and applied in our study but also the best system presented by the participants of the SemEval 2019 HatEval task in Spanish representing a considerable advance in the state-of-the-art. Within this study, we observed that traditional methods such as the ML classifiers explored are good estimators for the detection of misogyny and xenophobia in Spanish tweets.

The second study is among the first in the NLP community to investigate the new era of Transformer-based models for HS recognition in Spanish. In particular, we compared the performance between traditional methods and Transformer language models. Moreover, we also compared multilingual and monolingual Transformer language models trained specifically on Spanish to observe the language dependence of this problem. The results obtained show that the Transformer language models outperform traditional methods and the monolingual model (BETO) by a large margin. In this study, we also observed that the monolingual model (BETO) outperformed the multilingual models (mBERT and XLM) for HS detection, presumably because it has been trained specifically for Spanish and the vocabulary coverage of the dataset is higher than that of the multilingual ones. Therefore, this study shows, on the one hand, the remarkable capability of the new era of Transformer language models for the offensive language detection task as well as the importance of creating appropiate resources and models for Spanish, as the monolingual model BETO is able to more accurately modulate the vocabulary and specific expressions of the language than the multilingual models explored in this study.

An in-depth examination of the results obtained from both research studies shows that Spanish people use a varied vocabulary to refer to hatred. In particular, we observed some expressions that include hate terms and we found that sometimes, when they are combined with other terms, the sense changes completely. For example, the word *puta* (bitch) can be used in colloquial phrases or expressions with very positive polarity: *de*

puta madre (fantastic); and in other cases, aimed at men with negative polarity: *hijo de puta* (whoreson). For these reasons, we notice that HS detection in Spanish should be treated differently from English, since Spanish is a language with more semantic and morphological richness and complexity. Indeed, it is important to study its syntactic and semantics in order to develop accurate systems in this language and not to use systems adapted from other languages. In the error analysis conducted we have identified important challenges that the NLP models faced while identifying HS in Spanish. These include but are not limited to the following: the lack of context, the presence of offensive words not learned during the training procedure, the identification of the reference to other people's utterances, the use of insults with a positive connotation, the use of swear words in sentences with positive polarity, the use of rhetorical figures such as mockery, sarcasm, and irony, and the implicit expression of offensiveness.

In the following chapters, we rely on the main challenges observed in the preliminary studies while identifying offensiveness to continue the research of this doctoral thesis. First, we generate appropriate resources for offensive language detection in Spanish to overcome the great scarcity identified of linguistic resources labeled with offensiveness for this language (Chapter 4: *"Resource generation"*). Then, we propose a new solution to address the offensive language detection task, which is the main approach proposed in this doctoral thesis (Chapter 5: *"Combining linguistic phenomena through a multi-task approach"*). Specifically, due to the success observed by the Transformer language models in the second preliminary study, we aimed to continue in this direction by exploring these models. Moreover, based on some challenges observed by the automatic classifiers, we plan to provide more knowledge to the model by integrating different linguistic phenomena related to offensive language detection in order to help the model to generalize better on the task.

Chapter 4

Resource generation

This chapter describes the different linguistic resources generated in this doctoral thesis. First, the motivation for creating linguistic resources in Spanish for the detection of offensive language and emotion analysis is introduced. The methodology used to create these resources is then described. Following that, the lexical resource and the corpora that were constructed are presented. The chapter ends by discussing the main conclusions obtained in the course of this milestone that has been crucial to conduct the research in this doctoral thesis.

4.1 Motivation

Linguistic resources are fundamental for training and testing NLP-oriented systems. At the beginning of this doctoral thesis, we found that one of the major challenges in conducting the research was the lack of linguistic resources labeled with social phenomena such as emotions and offensive language, particularly in Spanish because most of them focused on English. However, the presence of different languages on the Web is growing every day, and therefore, it is important to invest efforts in the generation of resources focused on other languages. Moreover, both emotions and offensive language phenomena have a close relationship between the language and the context of its learning (social environment, cultural influences...). As a result, we realized that is essential to study these topics in different languages because there are significant cultural disparities in how emotions and offensiveness are presented.

Apart from being our mother tongue, Spanish is the world's second most spoken language and the third most utilized on the Web. Due to the high presence of this language on the Web as well as the importance of developing appropriate resources for implementing

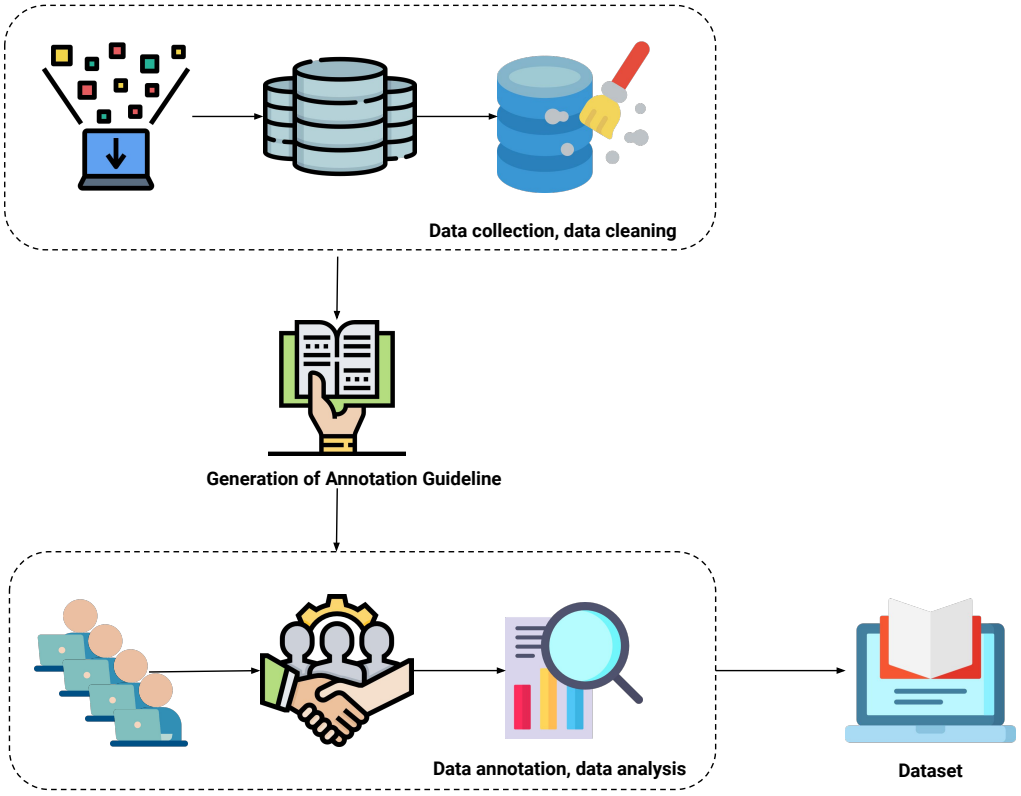

FIGURE 4.1: Resource creation process in NLP.

NLP-based systems to combat offensive language detection, one of the fundamental parts of the research in this doctoral thesis has been directed to the generation of linguistic resources for Spanish, specifically for emotion analysis and offensive language detection. Both lexical and corpus resources have been generated: a lexicon named SHARE, and three corpora named EmoEvent, OffendES, and OffendES_spans.

4.2 Methodology

In NLP, different steps are involved in the creation of a linguistic resource, from data collection with the corresponding cleaning to data annotation with the corresponding analysis. Figure 4.1 shows the different steps involved in the creation of a resource which are going to be described in detail in the next sections: data collection, data annotation, and data analysis. After these steps i.e., once the resource is created, an important step is the development of ML techniques to check its validity. For this reason, for each resource generated in this doctoral thesis, we have implemented a benchmark to evaluate the specific task and to provide the NLP community with preliminary results to compare future approaches.

In order to ensure the quality of the linguistic resources generated in this thesis, the following factors have been considered:

- **Large corpus size**. For each of the resources, the maximum amount of data possible has been retrieved. Training algorithms meant to perform specific tasks require large amounts of specialized datasets.
- **High-quality data**. When it comes to data within a corpus, high quality is critical. Because of the vast number of data necessary for a corpus, even little flaws in the training data might result in large-scale problems in the output of the machine learning system. In this step, the quality of the annotation of the dataset plays a crucial role, therefore, it is important to ensure that the annotations are reliable and that the inter-annotator agreement is as high as possible.
- **Curate data**. Data cleaning is also required for the creation and maintenance of a high-quality corpus. This step enables the detection and elimination of errors and duplicate data, resulting in a more accurate corpus for NLP.

4.2.1 Data collection

In NLP, the practice of acquiring accurate textual data from a range of relevant sources (surveys, websites, questionnaires, etc.) is known as data collection. In order to carry out this phase during this doctoral thesis, for each resource, we have followed the next steps: First, we identify the type of data needed to solve the specific problem or task, i.e., the source data as well as the specific metadata we want to retrieve. Next, we check the availability of data, i.e. how the information can be accessed for downloading and the quantity of information available. The last step is to use the identified mechanism to retrieve the data and decide how to structure and gather the information.

In recent years, social networks and messaging platforms have attracted the attention of users becoming an important part of our daily lives. Social media is used by billions of people around the world to share information and build connections. In 2021, it was estimated that in a single minute (60 seconds) around 695,000 stories were shared on Instagram; WhatsApp and Facebook Messenger users sent 69 million messages, and 500 hours of content were uploaded to YouTube. Therefore, nowadays we can easily retrieve a large amount of data generated by users in order to get a better understanding of the presence of different phenomena such as offensive language or the expressions of emotions, which are the research lines conducted in this thesis. According to a Spanish report in 2020 on the evolution of hate crimes in Spain[1], threats, insults, and

[1] `https://bit.ly/3SAokIm`

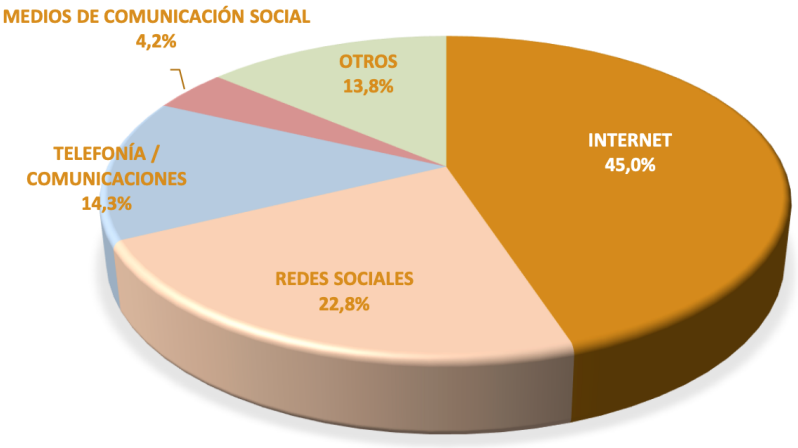

FIGURE 4.2: Means used for the commission of the hate crimes. Source: Spanish Ministry of the Interior.

discrimination are counted as the most repeated criminal acts, with the Internet (45.0%) and social media (22.8%) as the most widely used media to commit these actions (see Figure 4.2).

The accessibility of these data for download depends mainly on the social network. For instance, Twitter offers an API service called Twitter Search API[2] which is an HTTP-based RESTful API that returns responses encoded in JSON and can be used through an easy-to-use Python library to access the Twitter API: Tweetpy[3]. It allows users to download tweets and related metadata, such as the number of followers, likes, and retweets, by using a query in a specified language. Regarding Youtube, it also has an API that provides the ability to retrieve feeds related to videos, users, comments, and playlists, among others. For Instagram, the Graph API is available which be used to get and publish media content, manage comments and respond to comments on that content, identify media content with @mentions made by other Instagram users, search for media content with hashtags, and fetch results and basic metadata from other Instagram companies and creators. These Twitter, YouTube, and Instagram APIs need users to register as developers in order to gain access to the tokens and secret keys that allow data download.

4.2.2 Data annotation

The first step in the process of data annotation is to design the task through an annotation guideline created by experts. This guideline should be carefully prepared to give instructions on the task to the annotators. If these instructions are not clear, it may

[2] `https://developer.twitter.com/`
[3] `https://www.tweepy.org/`

Values	Interpretation
< 0	Poor agreement
0.00 to 0.20	Slight agreement
0.21 to 0.40	Fair agreement
0.41 to 0.60	Moderate agreement
0.61 to 0.80	Substantial agreement
0.81 to 1.00	Almost perfect agreement

TABLE 4.1: Interpretation of Cohen's kappa.

affect the performance of the annotators and thus the quality of the final resource. For a classification problem, an annotation guideline should include the following elements: the explanation of the task, the task's predefined categories and definitions, real-world examples of instances for each class, potential conflicting annotation scenarios, and frequently asked questions with the corresponding answers.

The second step consists of recruiting the annotators which are going to be involved in the task. A way to recruit annotators is through crowdsourcing platforms. One of the most popular crowdsourcing platforms is Amazon Mechanical Turk (MTurk). It is a forum where applicants post jobs as Human Intelligence Tasks (HITs). Workers complete HITs in exchange for compensation. It is possible to write, test, and publish the HIT using Mechanical Turk's isolated developer environment, MTurk APIs, and AWS SDKs. Customers who complete HITs are called **workers** and customers who publish these tasks are called **requesters**. Requesters can use the MTurk Web user interface to submit the task. The annotation provided by a worker for a HIT is called an **assignment**. It is also possible to indicate any additional requirements workers must meet to work on the task such as demographic, social, and professional features, among others.

Finally, after the annotation process, the agreement evaluation is performed. The Inter-Annotator Agreement (IAA) between the annotators is usually measured with Cohen's kappa metric [118], one of the most popular metrics used that expresses the level of agreement between annotators on a classification problem. It is defined in Equation 4.1, where p_0 is the empirical probability of agreement on any sample's label (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels at random. A per-annotator empirical prior over the class labels is used to estimate p_e [119]. This assessment indicates to us the quality of the agreement in the resource according to different values which are shown in Table 4.1.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{4.1}$$

4.2.3 Data analysis

After the data collection and data annotation, an in-depth study of a range of statistics from the created resource is required to get insight into how the specific task related to offensive language or emotion analysis is represented in the resource. These include, but are not limited to, the following:

- Number of posts by category.
- Post distribution by a specific characteristic involved in the resource (e.g., event, user profile, social media platform).
- In case of a multilingual dataset, number of posts per language.
- Part-of-speech tagging.
- Comments length.

It should be noted that depending on the specific corpus or lexicon, other statistics are performed in addition to these general statistics mentioned above.

In the following, we are going to describe in detail the different resources generated within the framework of this doctoral thesis.

4.3 SHARE

The generation of the SHARE lexicon will be detailed in this section. In particular, the phases outlined in the methodology section (data collection, data annotation, and lexical analysis) will be thoroughly described.

SHARE (Spanish HARmful Expressions) is a lexical resource composed of insults and offensive expressions manually labeled by 5 annotators. To collect the potentially offensive words and expressions provided by the Spanish speakers we used the Fiero chatbot developed by Botella-Gil et al. [120]. SHARE is available upon request to the authors.

The number of insults used in offensive comments can be unlimited depending on the imagination of the speakers, fashions, the influence of other languages, or the geographical context. Thus, although the Royal Spanish Academy (RAE: Real Academia Española) includes in its current dictionary a large number of insults such as *merluzo* (hake) or *ceporro* (dimwit), the richness of the language allows the creation of new words through composition. Spanish emerges as a great inventor of insults due to the

Unigrams	Bigrams	Trigrams	n-grams (N > 3)
91,005 (55.33%)	16,613 (10.10%)	14,649 (8.92%)	42,200 (25.65%)

TABLE 4.2: Total of n-grams in the data collection.

continuous evolution of the language and the emergence of new grammatical forms of verbal violence that are not included in the RAE [121]. For instance, the predilection for creating insults based on the word *cara* (face) is an example of this, which reaches current words such as *caranchoa* (anchovy face), passing through more subtle uses like *carajaula* (cage face). Furthermore, these insults can be formed by consecutive words, such as *chupa cabras* (suck goats) and *feo de mierda* (ugly shit).

The nature of some languages such as Spanish makes large-scale offensive lexicon development a difficult challenge. Since manual development is very costly and time-consuming, automatic and collaborative construction of computational lexical resources are real alternatives [122]. Moreover, lexical resources, such as lexicons are considered necessary to improve the performance of NER systems and interpretability tasks [123–125]. In making pre-trained models transparent and interpretable, it is often necessary to identify features that contribute significantly to a prediction.

4.3.1 Data collection

In order to collect the offensive terms present in SHARE, we used the virtual assistant in Telegram named Fiero [120]. Fiero was developed for encouraging users to insult in a humorous and sarcastic way with the aim of collecting insults and vulgar expressions from Spanish speakers. This tool was released in July 2019 and in 2020 it became more popular with significantly higher interaction due to the great diffusion and repercussion of Fiero in the radio, press and national television media.

A total of 164,467 comments were collected from 2019 to 2021. In this period, we obtained the number of comments shown in Table 4.2. 122,267 are composed of one, two, and three words (unigrams, bigrams, and trigrams, respectively). The remaining 42,200 are comments consisting of more than three words. We observed that more than half of the comments are composed of one term (unigrams). Table 4.3 shows the distribution of comments according to the gender and age of the users who interacted with Fiero. It can be seen that the male population over 18 years interacted the most, collecting a total of 95,513 comments. The younger population (<18) participates to a lesser extent, obtaining a total of 17,037 comments compared to 147,430 comments obtained by users over 18 years old.

Gender	Age	Comments
Female	>18	51,917
	<18	5,922
Male	>18	95,513
	<18	11,115
Total		164,467

TABLE 4.3: Total of n-grams obtained according to gender and age in Fiero.

Unigrams	Bigrams	Trigrams
11,936	6,930	7,765

TABLE 4.4: N-grams distributions in comments after preprocessing.

After the data collection, we accomplished different pre-processing steps by applying NLP-based automated techniques (both regular expressions and using the Python emoji library[4]):

- Comments were normalized to lowercase.
- Emojis were removed.
- Comments containing one only character, URL, punctuation marks, numbers, and consonants were deleted.
- Onomatopoeias such as haha, hehe, jaja, jeje including repeated characters and words that are part of the dialogue but not offensive (e.g., *hola* (hello), *adios* (goodbye), *sí* (yes), *seguro* (sure), *no*, *de acuerdo* (ok), *hola* (hello)) were removed.
- Elongated words and repeated characters were reduced. For instance, *toonnnto* (sssiilly) was replaced with *tonto* (silly).
- Comments longer than three words were deleted. We selected unigrams, bigrams, and trigrams to retrieve insults and expressions since consider that n-grams containing more than three words are part of comments involving a conversation.
- Duplicate comments were removed.

After the preprocessing phase, we obtained a total of 26,631 comments. Table 4.4 shows the distribution of these comments, 11,936 are unigrams, 6,930 bigrams and 7,765 trigrams.

[4] `https://pypi.org/project/emoji/`

Agreement			
Unigrams	Bigrams	Trigrams	All
0.6369	0.8183	0.8131	0.7881

TABLE 4.5: Kappa coefficient for inter-annotator agreement.

4.3.2 Data annotation

The final collected terms were labeled by five annotators. Specifically, we defined the following rules to annotate a term/expression as offensive or non-offensive:

- A comment is considered *offensive* when it contains some form of unacceptable language (profanity or bad words) or a targeted offense, which may be direct or indirect. This category includes insults, threats, and messages containing profane language or profanity. The message may be directed at an individual, at a group of people who share common characteristics, or at others (organization, situation, event, issue, or place) [126].
- Comments that contain the verb in front of a negative word, such as *eres idiota* (you are an idiot), are classified as non-offensive because we look only for bad words or expressions.
- Comments consisting of two or more consecutive offensive words are labeled as offensive, e.g. *idiota de mierda* (dumb shit).
- As a general rule, food and animal names are considered not offensive. However, there are some words in these contexts that are commonly used to offend. Therefore, we consider *perro/a, zorra, cerdo/a* (dog, fox, pig) as offensive.

Once the rules were defined, five annotators labeled a subset of the comments in order to compute the agreement. Specifically, each annotator labeled a total of 4,000 terms (2,000 unigrams, 1,000 bigrams, and 1,000 trigrams). After the first annotation, we computed the Cohen's kappa coefficient [118] to determine the agreement between the annotators. These results can be seen in Table 4.5. The results obtained with respect to the unigrams is 0.6369, which is considered according to [127] a substantial value. In the bigrams and trigrams, we obtain a value of near-perfect agreement, 0.8183 and 0.8131, respectively. With these results, we observed that comments composed of two or three words are easier to categorize as offensive than those consisting of only one word.

After the first annotation and examination that the agreement results obtained were favorable, each annotator labeled 4,927 new comments (2,187 unigrams, 1,286 bigrams,

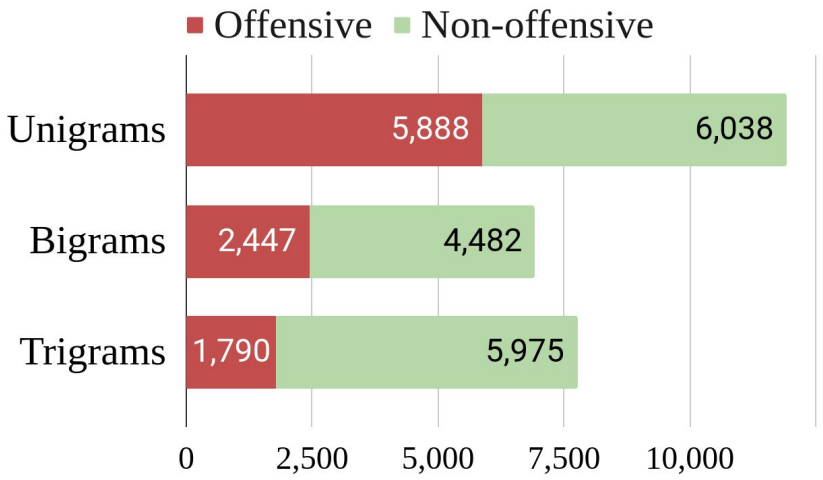

FIGURE 4.3: Distribution of the categories annotated according to n-grams selected.

and 1,453 trigrams), and one of the annotators also labeled an unigram to complete the total of 26,631 labeled comments.

4.3.3 Lexicon analysis

In order to perform a statistical analysis of the lexical resource developed, we analyzed the number of offensive and non-offensive terms and the distribution of n-grams labeled as offensive.

Figure 4.3 shows the distribution of the labeled categories according to the different n-grams taken into account, i.e. the number of offensive and non-offensive unigrams, bigrams, and trigrams. As we can see, the number of offensive (5,888) and non-offensive (6,038) unigrams are similar. When we analyze the bigrams, we can see that the number of non-offensive grows significantly to 4,482, almost double the number of offensive bigrams (2,447). Finally, we found 1,790 trigrams in the resource labeled as offensive and 5,975 in the non-offensive category. In total, SHARE is composed of 10,125 offensive expressions distributed as shown in Figure 4.4. As we can observe, the number of offensive unigrams represents 58.2% of the resource, which means that more than half of the resource is composed of a single offensive word. The remaining n-grams of the resource are covered by the offensive bigrams and trigrams, 24.2% and 17.7% respectively. These data were obtained taking into account that there was no overlap between unigrams, bigrams, and trigrams. For instance, in the trigram *hijo de puta* (son of a bitch), the word *puta* (bitch) is not considered as an unigram.

As far as we know, there are available two resources with Spanish offensive terms, the lexicons composed of 502 terms developed also in this doctoral thesis (see Chapter

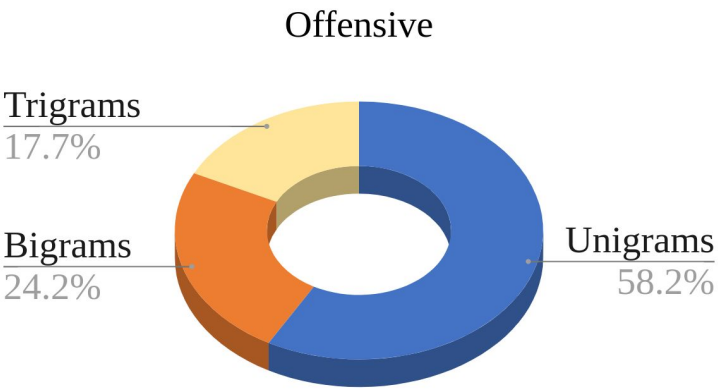

FIGURE 4.4: Distribution of n-grams labeled as offensive.

3: *"Preliminary research on offensive language detection"*) [94] and the multilingual HurtLex resource consisting of 2,933 unigrams [47]. We compare them with the SHARE resource in order to observe the difference in terms of size. The SHARE resource exceeds 9,623 insults to the lexicons built by Plaza-del-Arco et al. [94] and 7,192 terms to the HurtLex resource. In addition, we checked how many terms match with SHARE, finding that the lexicons built by Plaza-del-Arco et al. [94] contain 272 and HurtLex contains 247 matching SHARE terms. In summary, our lexicon offers a large number of offensive terms in the form of insults and expressions commonly used by Spanish speakers.

SHARE is also used to automatically annotate offensive entities in the OffendES corpus generated in this thesis. We also explore the usefulness of the lexicon as an interpretability tool for offensive comments by comparing it with a BERT-based fine-tuning model. These contributions are going to explained in Section 4.6: OffendES_spans.

The three corpora (EmoEvent, OffendES, OffendES_span) generated in this doctoral thesis will be detailed in the next sections. The steps involved in the creation of each dataset will be discussed in detail. Furthermore, for each dataset, the experiments conducted to validate the corpus and the preliminary results to the NLP community are presented.

4.4 EmoEvent

In recent years, emotion detection in text has become more popular due to its potential applications in fields such as psychology, marketing, political science, and artificial intelligence, among others. While opinion mining is a well-established task with many standard datasets and well-defined methodologies, emotion mining has received less attention due to its complexity.

In this section, we present a multilingual emotion dataset based on different events that took place in April 2019. We collected tweets from the Twitter platform. Then one of seven emotions, six of Ekman's basic emotions plus the "neutral or other emotions", was labeled on each tweet by 3 Amazon MTurkers. A total of 8,409 tweets in Spanish and 7,303 tweets in English were labeled. In addition, each tweet was also labeled as offensive or non-offensive. We report some linguistic statistics about the dataset in order to observe the difference between English and Spanish speakers when they express emotions related to the same events. Moreover, in order to validate the effectiveness of the dataset, we also propose a machine learning approach for automatically detecting emotions in tweets for both languages, English and Spanish.

EmoEvent is publicly available on the Hugging Face Dataset Hub: `https://huggingface.co/datasets/fmplaza/EmoEvent`.

4.4.1 Data collection

Our goal in collecting affective tweets is to explore great relevant events in a specific time frame on Twitter. In order to accomplish this, we focused on trending topic hashtags. Trending topics are the most used keywords during a given period of time on Twitter. It is a concept related to trends and topics, that everyone is talking about at any given time. In order to retrieve tweets for each event, we selected the trending topic that may contain affective content. In particular, we choose the following events that occurred during April 2019:

1. **Notre Dame Cathedral Fire**. On 15 April 2019, a structure fire broke out beneath the roof of Notre Dame Cathedral in Paris.
2. **Greta Thunberg**. She founded the movement "Fridays for Future". It refers to how she strikes every Friday to protest the lack of effective climate legislation on a governmental level. Students throughout Europe now regularly strike on Fridays.
3. **World book day**, is an annual event organized by the United Nations Educational, Scientific and Cultural Organization (UNESCO) to promote reading, publishing, and copyright. It is marked on April 23, the day of William Shakespeare's birth.
4. **Spain Election 2019**. The 2019 Spanish general elections were held on Sunday, November 10, 2019 to elect the XIII Cortes Generales of the President of Spain.
5. **Venezuela's institutional crisis**. A crisis concerning who is the legitimate President of Venezuela has been underway since January 10th of 2019, with the nation and the world divided in support of Nicolás Maduro or Juan Guaidó.

Event	*Hashtag (SP)*	*# of instances (SP)*	*Hashtag (EN)*	*# of instances (EN)*
Notre Dame	#NotreDameEnLlamas	24,539	#NotreDameCathedralFire	11,319
Greta Thunberg	#GretaThunberg	1,046	#GretaThunberg	1,510
World book day	#diadellibro	8,654	#worldbookday	17,681
Spain Election	#EleccionesGenerales28A	4,283	#SpainElection	493
Venezuela	#Venezuela	5,267	#Venezuela	5,248
Game of Thrones	#JuegoDeTronos	5,646	#GameOfThrones	9,389
La Liga	#Laliga	1,882	#Laliga	1,295
UCL	#ChampionsLeague	6,900	#ChampionsLeague	6,199

TABLE 4.6: Hashtags used to retrieve the tweets for each event and the total number of tweets retrieved in English (EN) and Spanish (SP).

6. **Game of Thrones**. This is an American fantasy drama television series. It is one of the most popular series in the world today. The last season premiered in April 2019.

7. **Campeonato Nacional de Liga de Primera Division (La Liga)** is the men's top professional football division of the Spanish football league system.

8. **The UEFA Champions League (UCL)** is an annual club football competition organized by the Union of European Football Associations (UEFA) and contested by top-division European clubs, deciding the best team in Europe.

We found these events very interesting because they belong to different domains such as entertainment (Game of Thrones, La Liga, UCL), catastrophes or incidents (Notre Dame Cathedral Fire), political (Venezuela's institutional crisis, Spain Election), global commemoration (World book day) and global strikes (Fridays for Future). Therefore, we were able to find a variety of emotions in the users who gave their opinions on these events.

Hashtag-based search on the Twitter search API. Trending topics are accompanied by hashtags that allow us to easily find all the tweets and conversations by users around that topic. In order to download the tweets, we used an easy-to-use Python library to access the Twitter API: Tweetpy[5]. It allows us to download messages using a query in a specific language. In our case, we chose as a query the trending topic hashtag associated with each event in English and Spanish, as can be seen in Table 4.6. For each tweet we obtained the following twitter metadata: *id*, *date*, *language*, *location*, *text*, *source*, *followers* and *friends*. We discarded tweets that had less than four words and tweets with very bad spelling. For this, we used a Python spell checker called hunspell[6] which contains a dictionary for English and Spanish. Also, we removed tweets with the prefix "Rt", "RT", and "rt", which indicate that the messages that follow are re-tweets (re-postings of tweets sent earlier by another user).

[5] `https://www.tweepy.org/`

[6] `https://pypi.org/project/hunspell/`

Post selection

One of the most commonly used techniques for choosing tweets from a dataset is random selection. However, the problem with this method in our case is that we can obtain many non-affective tweets. Since our goal was to get a dataset mainly labeled with emotions, we followed another strategy for selecting tweets, that of performing a linguistic analysis. It is based on extracting affective features from tweets using the Linguistic Inquiry and Word Count (LIWC) resource [128]. This is a popular content analysis technique that counts the occurrences of words according to pre-defined psychological and linguistic categories. The LIWC categories are grouped under four main dimensions: Linguistic Dimensions (e.g., word count, pronouns, negations, numbers); Psychological Processes (e.g., positive or negative emotions); the Relativity dimension describes physical or temporal information (e.g., time and space); and Personal Concerns (e.g., occupation, leisure activities). LIWC analysis has been successfully applied to a wide range of data, including determining the linguistic characteristics of emotion, personality, gender, and genre (Hancock, et al. 2007; Nowson, et al. 2005). Indeed, this resource is available in English and Spanish. Relying on this resource we focus on the dimension of psychological processes, extracting the following features:

- **Number of affective tweets**. We consider that a tweet is affective if it contains one or more words found in the affective category of LIWC. Otherwise, we assume that the tweet is not affective.
- **Number of positive tweets**. We consider that a tweet is positive if it contains more positive words than negative words. We checked the presence of positive words in the tweets by taking into account the positive category of LIWC.
- **Number of negative tweets**. We consider that a tweet is negative if it contains more negative words than positive words. We checked the presence of negative words in the tweets by considering the negative category of LIWC.

In order to gain a better understanding of the presence of emotion in tweets, we followed a method for calculating a score associated with a given class, as a measure of saliency for the given class inside the tweets collection.

We define the class coverage in the tweets corpus T as the percentage of tweets from T belonging to class C:

$$Coverage_T(C_1) = \frac{\sum_{T_i \in C} Tweets}{Size_T} \tag{4.2}$$

Event	*Affective Class*		*Positive Class*	
	SP	EN	SP	EN
Notre Dame	1.37	2.45	0.71	1.43
Greta Thunberg	0.86	1.64	1.31	2.46
World Book Day	1.36	2.25	6.85	12.51
Spain Election	0.92	2.01	1.59	5
Venezuela	1.47	1.44	0.94	1.12
Game of Thrones	0.88	1.53	1.12	1.29
La Liga	0.54	1.27	2.11	10.71
UCL	0.75	1.13	1.93	3.24

TABLE 4.7: Prevalent of each class in the different events.

The prevalence score of class C in the tweets corpus T is then defined as the ratio between the coverage of one class in the corpus T with respect to the coverage of the other class in corpus T.

$$Prevalence_T(C_1) = \frac{Coverage_T(C_1)}{Coverage_T(C_2)} \tag{4.3}$$

A prevalence score close to 1 indicates a similar distribution of the tweets between class C_1 and class C_2 in corpus. Instead, a score significantly higher than 1 indicates that class C_1 is prevalent in the corpus. Finally, a score significantly lower than 1 indicates that the class C_2 is dominant in the corpus.

In Table 4.7 we can see the prevalence of the affective and positive classes for the different events in Spanish and English. Interestingly, in both languages, the top events where the positive class is prevalent are the same: world book day, La Liga and the UCL. However, for the affective class, there are more differences between the two languages. For English, we find that there is a greater prevalence in the affective class than for Spanish. This means that for these events English speakers express more emotions in tweets than Spanish speakers.

After analyzing the affective and non-affective tweets, we decided to randomly select 1,000 affective tweets and 200 non-affective tweets for each language and event in order to perform the annotation. The final dataset distribution is shown in Table 4.9.

4.4.2 Data annotation

Annotations were obtained via the MTurk platform. This is a powerful vehicle for getting tasks done quickly and efficiently. As a requirement for annotators, we indicate that they must be located in Spain (ES) to label the Spanish dataset and the United States (US)

Emotion	*SP*	*EN*
anger	44.18	19.52
sadness	55.55	38.81
joy	41.10	36.68
disgust	18.61	20.96
fear	29.70	10.08
surprise	17.00	13.22
offensive	54.67	22.15
other	34.78	18.76

TABLE 4.8: Kappa coefficient for inter-annotator agreement.

to label the English dataset. We created HITs for each of the tweets corresponding to the events specified in Table 4.6. Each HIT had two questions, answered by three different workers. The first question is designed to label the main emotion conveyed by the tweet (*anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise* or *others*), the second one to determine whether the tweet contains offensive language or not.

The annotation guideline designed to label the posts is shown in Appendix A.

After the three workers had completed labeling the dataset, we decided the final tweet label based on their annotations in the following way: If two or three annotators agree on the same emotion, we label the tweet with that emotion. Otherwise, we label the tweet as *other*.

Inter-Annotator Agreement. In order to analyze how often the annotators agreed with each other, we conducted inter-tagger agreement studies for each of the eight categories. For this, we use the Cohen's kappa coefficient and the values are shown in Table 4.8. In order to measure the level of agreement among the three annotators, we measured the agreement between each annotator and the average of the remaining two annotators. As we can see in Table 4.8 the agreement between the Spanish annotators for each emotion is higher than the one obtained by the English annotators. It can also be noted that the most difficult emotions to label by the annotators are *disgust*, *fear* and *surprise* for both languages.

4.4.3 Corpus analysis

In this section, we highlight some statistics regarding the multilingual emotion dataset. These statistics refer to the number of tweets by event, hashtags, emojis, and part-of-speech, among others.

Table 4.9 shows the number of tweets selected by event and language, the average length of the tweets, the number of emojis, and the number of unique hashtags found in the

Event	*# of tweets*		*Avg. tweet length*		*# of emojis*		*# of unique hashtags*	
	SP	EN	ES	EN	SP	EN	SP	EN
Notre Dame	1,200	1,200	26.57	26.98	432	242	397	942
Greta Thunberg	630	742	24.91	27.61	279	154	750	1,036
World Book Day	1,200	1,200	23.93	23.83	916	649	827	1,131
Spain Election	1,200	207	20.89	24.67	355	37	373	185
Venezuela	1,200	1,200	24.16	25.16	238	163	681	735
Game of Thrones	1.200	1,200	19.86	21.80	579	565	372	343
La Liga	579	354	19.38	17.70	712	511	372	311
UCL	1,200	1,200	16.77	18.30	782	776	386	641
Total	8,409	7,303	22.06	23.26	4,293	3,097	4,158	5,324

TABLE 4.9: Number of tweets by event, average length of tweets, hashtags and emojis in the dataset.

Event	*joy*		*anger*		*fear*		*sadness*		*disgust*		*surprise*		*other*	
	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN
Notre Dame	59	148	153	78	2	20	660	234	34	218	27	41	265	461
Greta Thunberg	80	33	33	2	1	9	14	4	3	10	11	5	488	144
World Book Day	465	419	13	39	0	20	32	19	5	74	13	61	672	568
Spain Election	316	190	170	3	44	0	58	6	38	5	38	4	536	146
Venezuela	92	59	283	175	18	57	119	59	55	260	20	20	613	570
Game of Thrones	269	647	107	7	29	3	87	8	9	26	173	12	526	497
La Liga	184	177	23	30	0	28	10	7	1	98	17	6	344	396
UCL	350	366	75	58	2	14	29	79	16	74	45	86	683	523
Total	1,815	2,039	857	392	96	151	1,009	416	161	765	344	235	4,127	3,305

TABLE 4.10: Number of tweets by emotion and event in the dataset.

Event	*# of offensive tweets (SP)*	*# of offensive tweets (EN)*
Notre Dame	80	116
Greta Thunberg	6	20
World Book Day	17	24
Spain Election	146	4
Venezuela	184	150
Game of Thrones	165	122
La Liga	17	10
UCL	91	72
Total	706	518

TABLE 4.11: Number of offensive tweets in English (EN) and Spanish (SP) in the dataset.

dataset. It contains a total of 8,409 tweets for English and 7,303 for Spanish. It should be noted that Spanish users tend to use more emojis than English users to express their opinions on different events. However, hashtags are more used by English users.

The number of emotion tweets per incident is shown in Table 4.10, where we can determine which emotions are dominant for each one. World book day was the predominant event for the *joy* emotion. *Anger*, *disgust* and *fear* were more usual for the Venezuela

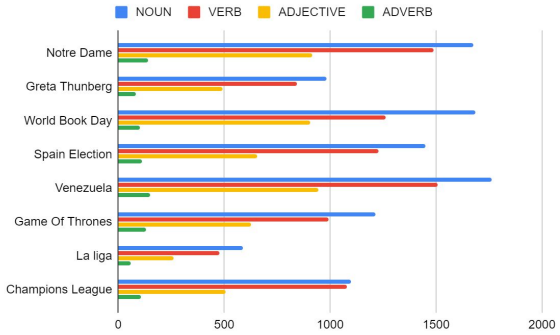

FIGURE 4.5: Part-of-speech tagging in the SP dataset.

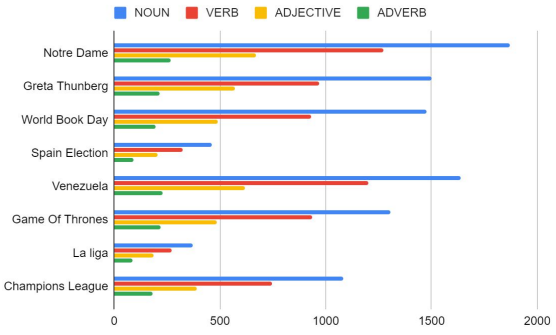

FIGURE 4.6: Part-of-speech tagging in the EN dataset.

situation. *Sadness* was the most frequent emotion in the case of the Notre Dame Cathedral Fire disaster. *Surprise* was more present at entertainment events such as Game of Thrones and UCL. It is necessary to emphasize that there are some emotions that are difficult to label by human annotators. For example, it can be observed that the number of tweets for *fear*, *disgust* and *surprise* are noticeably lower compared to others (*joy*, *sadness*, *anger*). In particular, *fear* and *surprise* are the most difficult emotions to label. This is because while instances of some emotions tend to be associated with exactly one valence (eg, *joy* is always associated with positive valence), instances of other emotions can be associated with differing valence (sometimes *surprise* or *fear* are associated with positive valence, while other times they are associated with negative valence) [129]. Therefore, an annotator can be confused to find an opinion that expresses *surprise* but also *joy*. In this case, most of the time the opinion is labeled by the annotator as *joy*.

Table 4.11 shows the number of offensive tweets per event in English and Spanish in the dataset. In general, there were few offensive tweets for each event. It is remarkable that in both languages the most offensive tweets were associated with the Venezuelan political incident.

The grammatical labeling for English and Spanish can be found in Figures 4.5 and 4.6. As can be seen, Spanish users tend to use more nouns, verbs, and adjectives to express their emotions. However, this is not the case with adverbs, which are more widely used by English users.

4.4.4 Experiments and results

In this section, we describe the different experiments we carried out to test the validity of the EmoEvent dataset generated. In particular, we trained a classifier based on the traditional ML paradigm, the Support Vector Machine (SVM).

Pre-processing. Pre-processing the data is the process of cleaning and preparing the text for classification. It is one of the most important steps because it should help improve the performance of the classifier and speed up the classification process. Online texts usually contain a great deal of noise and uninformative parts which increases the dimensionality of the problem and hence makes the classification more difficult. For this reason, we applied pre-processing techniques in order to prepare the data for the text classification. In particular, we preprocessed the tweets following these steps: The tweets were tokenized using NLTK TweetTokenizer[7] and all hashtags were removed.

Classification. Features in the context of text classification are the words, terms or phrases that express the opinion of the author. These have a greater impact on the orientation of the text. There are several ways to assess the importance of each feature by attaching a certain weight to it in the text. We use the most popular: The Term Frequency Inverse Document Frequency scheme (TF-IDF). Specifically, using this scheme each tweet is represented as a vector of unigrams. Machine learning techniques are popular in the classification task. For this reason, we decide to employ a machine learning algorithm in order to classify the tweets by emotions. In particular, we selected the Support Vector Machine (SVM). It is one of the most well-known classifiers since it has been shown to be highly effective and accurate in text categorization.

Results

In order to evaluate and compare the results obtained by our experiments we use the usual metrics in text classification: precision, recall, F_1-score, and accuracy.

We used 10-fold cross-validation to evaluate the machine learning classification approach. The results achieved with the SVM algorithm on the multilingual dataset are shown in Table 4.12. As can be seen, we achieved better results for Spanish (*Acc:* 0.64) than for English (*Acc:* 0.55). For both languages we obtained the best scores on *joy*, *sadness* and *other* labels. However, the other emotions (*anger*, *fear*, *disgust* and *surprise*) are not as easy to detect for our classifier and specifically for English. This may be because we have a lower number of tweets labeled with those emotions and also because they are complementary emotions. It means that, for instance, *anger* and *disgust* may occur at the same time. In fact, the annotators in the labeling process have found it difficult to discern between these two emotions. The same can happen with the *surprise* emotion. Finally, it is important to mention that while in the Spanish dataset we get a good score for the sad emotion (F_1*:* 0.70), this does not occur for the English dataset, where the score is noticeably lower (F_1*:* 0.46).

[7] `https://www.nltk.org/api/nltk.tokenize.html`

Language	*joy*			*sadness*			*anger*			*fear*			*disgust*			*surprise*			*other*			*macro-avg*			
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	*Acc*
SP	0.60	0.49	0.54	0.79	0.63	0.70	0.55	0.34	0.42	0.63	0.36	0.46	0.21	0.02	0.03	0.39	0.12	0.19	0.64	0.84	0.73	0.54	0.40	0.44	0.64
EN	0.59	0.60	0.6	0.62	0.36	0.46	0.33	0.10	0.16	0.35	0.04	0.07	0.38	0.21	0.27	0.16	0.02	0.04	0.54	0.73	0.62	0.42	0.29	0.32	0.55

TABLE 4.12: Results obtained from the multilingual dataset (10-fold cross-validation) with SVM. P: Precision, R: Recall.

4.5 OffendES

Another corpus generated in this doctoral thesis is OffendES, a new corpus in Spanish for offensive language research. To understand the rationale behind the design and generation of this corpus, certain contextual information may be useful. On the one hand, most of the corpus generated for offensive language research focused on English. In addition, most of them have been focused on Twitter data, despite the presence of offensive language on other platforms such as YouTube or Instagram, which are more widely used by young people. On the other hand, as mentioned above, dealing with offensive posts on social networks is a growing concern. Several platforms are clear on this issue, as can be read in rules and policies of Twitter[8], Instagram,[9] or YouTube[10]. Indeed, YouTube has disabled comments on videos and channels featuring children [130]. But this is a major concern not only for platform providers but for public administrations, in order to limit the possible side effects of harmful messaging to more vulnerable communities, like children or teenagers. With this in mind, the creation of OffendES aims to achieve the following long-term goals:

1. Early detection of offensive language use in social media on the Internet, with a special focus on young people.
2. Study of offensive language not only on Twitter but also on YouTube and Instagram.
3. Identifying improvements in protection systems for young people in social networks.
4. Studying the feasibility of automatic learning systems for offensive language in Spanish.
5. Creating a reference corpus for the study of language technologies applied to the classification of sexist language.

OffendES is publicly available on the Hugging Face Dataset Hub: `https://huggingface.co/datasets/fmplaza/offendes`.

4.5.1 Data collection

Instagram, YouTube, and Twitter are among the social media platforms most used by people ages from 18 to 24 [131]. These three have been selected as the main data

[8] `https://cutt.ly/1j5Eut0`
[9] `https://cutt.ly/yj5Eijc`
[10] `https://cutt.ly/kj5Eo2d`

Social network	Offensive terms	Non-offensive terms	Total
YouTube	19,449	184,414	203,863
Instagram	3,142	58,209	61,351
Twitter	1,197	18,728	19,925
Total	23,788	259,865	283,622

TABLE 4.13: Presence of offensive terms from lexicons in the retrieve comments.

sources. A total of 12 controversial influencers with a significant number of followers have been identified and their respective accounts in the three targeted social media platforms have been tracked. They are Spanish influencers from 24 to 35 years old and, six are men and six are women. The process for collecting comments consisted of two main steps. To collect the data, first, the last 50 posts by each influencer were obtained using the platform API. Then, an *ad hoc* web scraper was launched to extract user comments to each of the posts obtained (limited to 2,000 replies). This script uses scrolling through JavaScript code commands to retrieve further comments. In the case of YouTube, instead of the scraper, its API[11] has been used to retrieve comments.

Post selection

During two months (from February to March 2020), a total number of 283,622 comments were collected (see Table 4.13 for detailed information). The comments were then filtered according to two main constraints: the presence of potentially offensive language and lexical diversity.

To avoid the creation of a corpus with few or no offensive comments set, we labeled all the comments with flags determining whether the comment contained any of the words found in five different controlled lexicons [94]. All comments with potentially offensive language were selected (23,788 comments). We selected 60,000 comments to be labeled in the manual annotation phase. Therefore, we selected 36,212 comments without offensive terms. Applying lexical diversity measures proved to be an interesting approach to ensure a diverse set of comments. Therefore, we first attempted to include those comments that added the highest lexical diversity value to the growing set of collected comments. To that end, we applied the Measure of Lexical Textual Diversity (MTLD) [132], but the expected time to build the corpus with our implementation was unacceptable. Thus, we simply added those comments that produced the highest increase in the vocabulary size to the collection by iterating through all the comments and checking the amount of increase in vocabulary size comment by comment. At each iteration, the comment with the highest contribution of new vocabulary to the final collection was selected. This process was repeated until 60,000 comments were reached.

[11] `https://cutt.ly/JkrVSYv`

4.5.2 Data annotation

In order to establish the annotation schema, we followed those defined in [5, 23], while introducing some additional details that we consider important. Namely, we created a new category to include those posts with inappropriate language but no offense intended. For instance, the comment "eres la puta ama" (*you're the fucking boss*) contains inappropriate but non-offensive language and has a positive polarity. Then, we reformulated the definition of offensiveness to not include such posts.

The previous analysis led us to propose a definition of an offensive comment: one where language is used to commit an explicit or implicitly directed offense that may include insults, threats, profanity, or swearing. Based on this definition, we established the following categories:

- **Offensive, the target is a person (OFP)**. Offensive text targeting a specific individual.
- **Offensive, the target is a group of people or collective (OFG)**. Offensive text targeting a group of people belonging to the same ethnic group, gender or sexual orientation, political ideology, religious belief, or other common characteristics.
- **Offensive, the target is different from a person or a group (OFO)**. Offensive text where the target does not belong to any of the previous categories, e.g., an organization, an event, a place, an issue.
- **Non-offensive, but with expletive language (NOE)**. A text that contains£ rude words, blasphemes, or swear words but without the aim of offending, and usually with a positive connotation.
- **Non-offensive (NO)**. Text that is neither offensive nor contains expletive language.

The detailed annotation guideline designed to label the posts is shown in Appendix B.

The annotation of the collected data was performed via MTurk[12], which is a popular crowdsourcing platform. For the requirements of the human annotators, we selected the location as Spain and the time to five minutes due to the presence of some long comments from YouTube. Apart from releasing the annotation scheme with four examples of instances for each class, for the purpose of ensuring clear and concise documentation,

[12] `https://www.mturk.com/`

we also provided a list of instructions about rules, tips, and FAQs to try to solve any potential problems that could arise during the labeling process. Finally, to ensure the quality of the annotations, we used tracking comments.

We first conducted a round of trial annotation for both types of labeling, 4,500 and 1,500 instances with three and ten annotators, respectively. The goal of the trial annotation was *(i)* to identify any confusion in understanding the annotation schema, *(ii)* to estimate the average time to label the dataset, and *(iii)* to learn about the platform. The launch of these datasets was on September 24th, 2020, and it took two weeks to complete the annotation process on both sets. After analyzing the annotations, we observed through the comments of the annotators that the NOE and OFO classes were the most difficult to identify in the comments by the annotators. For this reason, we improved the definition of each class, providing examples as clear as possible to the annotators. The average agreement (kappa coefficient) grew from 36.85% for trial annotations up to 39.37% for final released comments. Yet, this level of agreement is lower than expected, which reflects the difficulty to discriminate among proposed classes.

Once the trial round was completed, the next step was to release the final dataset. A total of 54,023 instances were released in two subsets: 40,513 labeled by three annotators, and 13,510 labeled by ten annotators. The annotation took place from November 17, 2020 to January 2, 2021. As result, the three annotators subset covered 44,951 comments and the ten annotators subset 14,989 comments.

Post-processing. In order to check the reliability of the annotators, we analyzed their annotations in the tracking comments, i.e. those comments given as examples in the annotation guide. We observed that one of the annotators had over 60% of error rate in the tracking comments of both types of labeling, so we decided to remove their annotations since they could negatively affect the quality of the dataset. Sadly, this annotator was one of the most prolific, so the removal of his/her annotations resulted in a reduction of the three annotators' subset to a number of 44,951 comments.

4.5.3 Corpus analysis

Thus, the final dataset is released divided into two subsets: the three annotators subset (3-Ann), with 44,951 comments, and the ten annotators subset (10-Ann), with 14,989 comments. The former is intended for multi-class classification research and the latter for tackling multi-output regression problems. Only 38 comments belong to both subsets. Comments are compiled without processing, therefore, case, punctuation, and emojis are preserved. Every comment is associated with a social network platform (Instagram, Twitter, or YouTube) and directed to one of the 12 selected influencers as the target.

	3-Ann Subset				10-Ann Subset			
Influencer	*Instagram*	*Twitter*	*YouTube*	*Total*	*Instagram*	*Twitter*	*YouTube*	*Total*
dalas	3,558	1,454	6,813	*11,825 (26.3%)*	1,223	494	2,214	*3,931 (26.2%)*
soyunapringada	582	31	5,412	*6,025 (13.4%)*	172	7	1,745	*1,924 (12.8%)*
windygirk	466	487	3,756	*4,709 (10.5%)*	183	186	1,249	*1,618 (10.8%)*
javioliveira	276	130	3,890	*4,296 (9.6%)*	92	52	1,297	*1,441 (9.6%)*
wismichu	859	327	2,929	*4,115 (9.2%)*	318	101	1,014	*1,433 (9.6%)*
miare	508	167	2,749	*3,424 (7.6%)*	166	63	936	*1,165 (7.8%)*
wildhater	648	0	2,485	*3,133 (7.0%)*	204	0	843	*1,047 (7.0%)*
nauterplay	540	0	2,058	*2,598 (5.8%)*	180	0	685	*865 (5.8%)*
lauraescane	286	152	1,991	*2,429 (5.4%)*	107	50	633	*790 (5.3%)*
dulceida	226	0	1,400	*1,626 (3.6%)*	81	0	440	*521 (3.5%)*
jpelirrojo	69	0	582	*651 (1.4%)*	23	0	187	*210 (1.4%)*
nosoymia	107	13	0	*120 (0.3%)*	42	2	0	*44 (0.3%)*
Total	*8,125*	*2,761*	*34,065*	*44,951*	2,791	955	11,243	*14,989*
	(18.6%)	*(6.4%)*	*(75.0%)*	*(100.0%)*	*(18.1%)*	*(6.1%)*	*(75.8%)*	*(100.0%)*

TABLE 4.14: Comments per social media and influencer in the OffendES dataset.

Label	**3-Ann**	**10-Ann**
NO	26,425	9,715
OFP	4,102	2,362
NOE	2,470	1,414
None	11,529	1,283
OFG	425	215

TABLE 4.15: Comments per label in the OffendES dataset.

In Table 4.14, the amount of comments associated with each platform and influencer is depicted. Comments on *dalas'* posts are more frequent (over 26% in both subsets). YouTube is the platform where most of the comments were collected (about 75% for both subsets), followed by Instagram (over 18%). Comments from Twitter only represent just over 6% of the collection.

For both subsets, the final label is the majority class according to human annotators. For the subset labeled by ten annotators, the majority vote was set to five annotators. An additional *None* label was used when no agreement was reached between annotators. Table 4.15 shows the number of comments for each label on both subsets. Noticeably, the 10-Ann subset has a much lower percentage of *None* labels than the 3-Ann subset. The more annotators that were involved, the easier it was to decide the final label for a comment.

Table 4.16 shows statistics on comments length (i.e. the number of characters in the text). As expected, YouTube is the platform with the highest average length (about 190 for both subsets), with high variance; Twitter comments average length is lower (149 characters), with very small variance, and Instagram is the platform where comments tend to be the shortest (with an averaged length of 114).

(3-Ann subset)	**Average**	**Std. dev.**	**Min.**	**Max.**
YouTube	189	247	3	9,986
Twitter	149	75	4	413
Instagram	114	124	3	2,200
(10-Ann subset)	**Average**	**Std. dev.**	**Min.**	**Max.**
YouTube	191	277	4	9,812
Twitter	150	74	5	292
Instagram	113	115	3	1,631

TABLE 4.16: Statistics over comments length.

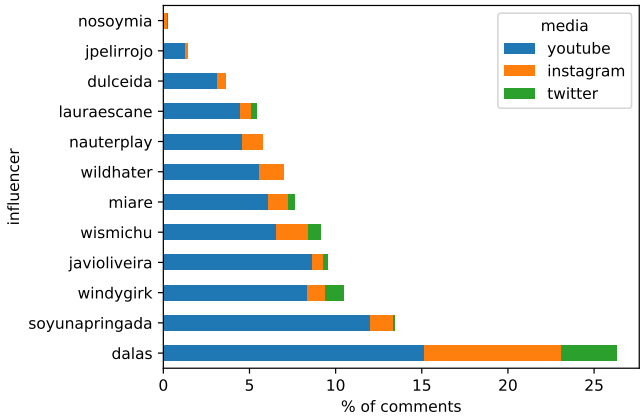

FIGURE 4.7: Comments distribution by influencer and social media platform in the 3-Ann subset.

Figure 4.7 shows the distribution of comments among influencers and social media platforms in the 3-Ann subset. YouTube is the most frequent platform, followed by Instagram. The influencer *dalas* is the target of more than a quarter of the total amount of comments. A similar distribution of comments is found in the 10-Ann subset.

An interesting analysis is to measure label frequency according to each influencer. Figure 4.8 shows the proportion of influencer-level labels and reflects the differences among these users as a target of offensive comments. In terms of gender, it can be seen that female influencers are subject to a greater number of offensive comments than male accounts. In particular, *soyunapringada*, *miare_love*, and *WindyGirk* are the accounts ranked with the most offensive comments. Regarding male influencers, accounts like *JaviOliveira* and *NauterPlay* contain more offense comments than accounts like *WildHater* and *JPelirrojo*. The profile of the influencer may define more controversy compared to others, or raise more negative emotions in their followers. Therefore, it could be interesting to consider the target profile as a source of information in offensive detection systems.

Inter-annotator agreement using the three annotators subset was measured with Cohen's kappa coefficient. The k value is 0.3579 (fair agreement), which is quite low and reflects

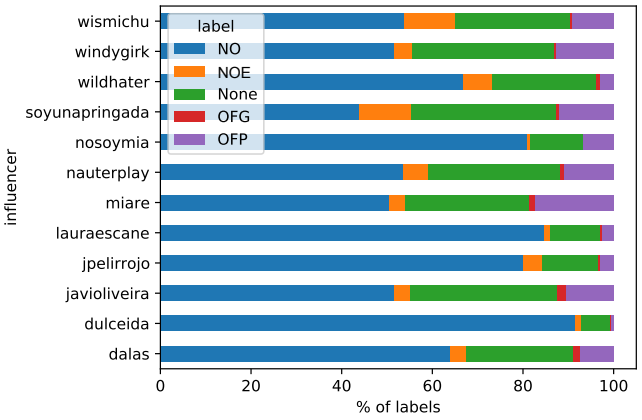

FIGURE 4.8: Distribution of labels per influencer in the OffendES dataset.

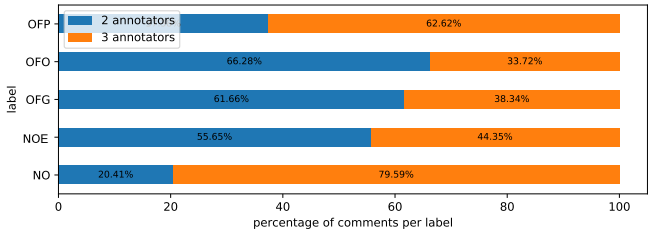

FIGURE 4.9: Percentage of consensus per label.

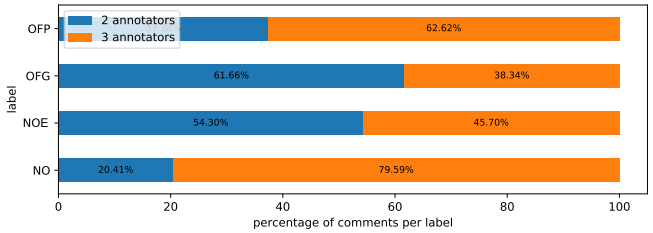

FIGURE 4.10: Percentage of consensus per label after including OFO label into NOE.

how difficult it is for humans to agree between the proposed categories. By analyzing annotations on tracking comments, we found that it was a common mistake to label a comment NOE or OFG when it should have been labeled OFO. Figure 4.9 shows the percentage of consensus per label in the subset of 3-Ann taking as consensus the majority vote (2-annotators agreement and 3-annotators agreement). As can be noticed, the label OFO exhibits the lowest consensus rate, with all three annotators only agreeing on 33.72% of the time. We found that many OFO comments were wrongly annotated with the NOE label and this could be reasonable since these offenses are not directly targeted at persons or groups, and they often consist of expletive expressions. Thus, we decided to merge them. After merging the OFO label into the NOE label, the kappa value increases slightly up to 0.3837. Figure 4.10 shows the final percentage of consensus per label after the merge of NOE and OFO labels.

		MTLD
Social network	Instagram	42.14
	Twitter	61.74
	YouTube	60.59
Label	NO	66.36
	NOE	26.41
	None	53.59
	OFG	53.19
	OFP	28.68

TABLE 4.17: Average values of measures of lexical textual comments diversity per social network and label.

Another feature we analyzed is the lexical diversity of comments. To this end, we use the MTLD metric already introduced, which allows us to get an insight into lexical variation and avoid biases due to different text lengths. Table 4.17 shows the average values for MTLD for comments over labels and platforms, respectively.

As can be noticed, offensive comments targeted at a person (OFP) have low lexical diversity, as well as for those with expletive language (NOE). When the comment is not offensive at all, the lexical diversity is clearly higher. Regarding social networks, we would expect the lowest value of diversity on Twitter, as it limits comment length. On the contrary, Twitter is the platform with the highest lexical diversity, followed by YouTube. Instagram is clearly much poorer in terms of the diversity of vocabulary used. These findings are worth exploring, as they could provide more understanding of how language is used across platforms and how it relates to harmful language use, or on the average profile of their communities. To understand MTLD values, we have to consider that a value of 50 is the average lexical diversity of texts for an average adult text (being 80 for academic writings).

4.5.4 Experiments and results

In order to establish a baseline for the OffendES corpus, we conducted experiments based on three different approaches: simple majority class model, lexicon-based model, and Transformer-based model.

Simple majority class model. Our simplest classifier assigns the majority class of the training set, i.e., the NO class, to each instance in the test set. This results in accuracy values of 58.78% and 64.85% respectively for 3-Ann and 10-Ann subsets.

Lexicon-based model. We also developed a lexicon-based approach using the lexical resources described in Section 4.5.1. In this approach, we only consider a binary

classification scenario: whether the comment is offensive or not. For the 3-Ann subset, we obtained 67.13% of accuracy, 21.27% precision, 83.78% recall, and 33.93% F_1. For the 10-Ann subset, the values of accuracy, precision, recall, and F_1 were, respectively, 71.45%, 35.59%, and 81.60%, 49.56%.

Transformer-based model. Finally, we experimented with a Spanish pre-trained BERT model called BETO [112] which has shown promising results in offensive language detection tasks [95]. In order to evaluate the model, we sampled from the collection two different sets, for training and evaluation. Measures used to report performance are precision, recall, and F_1-score at class level, and macro and weighted average of these metrics. For the multi-output regression task, since we are not dealing with a multi-class scenario, we used one of the most preferred metrics for regression tasks, the mean MSE, a risk metric corresponding to the expected value of the squared (quadratic) error or loss (see Equation 4.4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{4.4}$$

Multi-class classification

This experiment is performed on the 3-Ann subset. All entries labeled as *None* were discarded (as no final label was assigned to these comments). The set was split into training (95%) and evaluation (5%) partitions, resulting in 30,079 comments in the training set and 3,343 in the evaluation set. Transformers [133] library by Huggingface[13] was used to build the BERT network and the tokenizer from available BETO models (uncased variant).

A sequence classifier was implemented for this multi-class task, with a final linear layer with four outputs (the logits for each possible label). Training the model took 2 hours and 26 minutes.

After seven training epochs, the model was evaluated against the evaluation partition. The results obtained are depicted in Table 4.18.

Binary classification with BETO

Same configuration as the previous model, but using non-weighted cross-entropy as loss function during training. Classes have been merged into two classes as follows: *Non-offensive*, which comprises labels NO and NOE, and *Offensive*, combining OFP and OFG labels. This results in 28,895 non-offensive comments and 4,527 offensive comments. The results obtained are depicted in Table 4.19.

[13] `https://huggingface.co`

Class	P (%)	R (%)	F_1 (%)
NO	95.24	87.88	91.42
NOE	57.86	79.31	66.91
OFP	57.48	68.87	62.66
OFG	30.00	52.17	38.10
macro	60.15	72.06	64.77
weighted	86.96	84.39	85.33

TABLE 4.18: Multiclass experiment results. P: Precision, R: Recall.

Class	P (%)	R (%)	F_1 (%)
Non-offensive	92.79	95.14	93.95
Offensive	68.06	58.33	62.82
macro	80.42	76.74	78.39
weighted	89.06	89.59	89.26

TABLE 4.19: Binary classification experiment results. P: Precision, R: Recall.

Multi-output regression with BETO

For every sample, a vector of probabilities is computed by counting the number of annotators that selected each label and dividing by the number of annotators. This provides an estimate of the confidence of each label to be assigned to the comment. Training the model took 48 minutes.

The 10-Ann dataset was split into training and validation partitions. After training for seven epochs over a partition of 13,020 samples, the model was evaluated against a partition of 685 test samples, obtaining an MSE of 0.0241.

4.6 OffendES_spans

The development of the SHARE resource allows not only the detection of Spanish offensive texts but also the automatic annotation of offensive entities in corpora for NER tasks. For this aim, we leverage OffendES, to demonstrate the validity of the SHARE resource for NER. OffendES_spans is available upon request to the authors.

4.6.1 Data annotation

We automatically annotated the OffendES corpus described in Section 4.5 with the terms included in SHARE (see Section 4.3: *SHARE*), and we named this new resource as

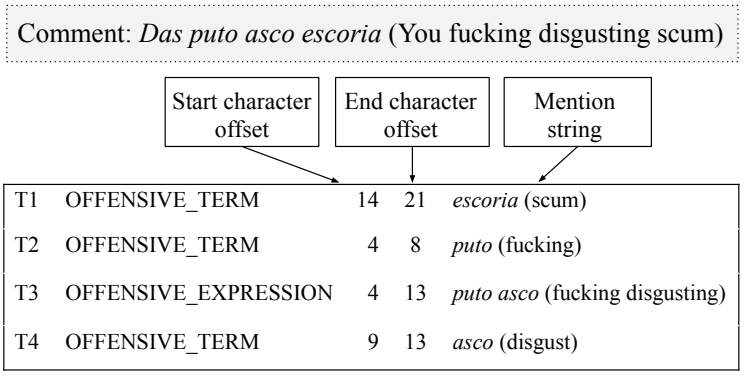

FIGURE 4.11: An example of an annotation file in OffendES_spans corpus.

OffendES_spans. This strategy involves performing different processing steps to properly match the comments in the corpus with the offensive terms.

The gold standard OffendES_spans corpus has been distributed in CSV format with different fields such as comment, social network, influencer, and label, among others. The annotations of offensive terms are included in a separate document (ANN file), with the same name as the ID of the comments.

Two types of entities can be found within the ANN files: OFFENSIVE_TERM, which refers to offensive unigrams, and OFFENSIVE_EXPRESSION, to label entities composed of more than one word (i.e. bigrams and trigrams). Every line of the ANN file contains the mention string of the annotation, its start character offset, and its end character offset, which uniquely locates the mention in the text comment. See Figure 4.11 for an example of the tab-separated file with the annotation information.

4.6.2 Corpus analysis

After the automatic annotation, we analyzed the offensive terms included in the OffendES comments, i.e, the spans annotated in the OffendES_spans corpus. The 12 most frequent terms annotated are presented in Table 4.20. As can be seen, the most commonly used offensive terms are *mierda* (shit), *puto* (whore) and *puta* (bitch). Related to bigrams and trigrams, the most frequent ones in the corpus are *puta madre* (fucking mother), *mala persona* (bad person), and *cacho de mierda* (piece of shit). Other terms such as *ignorante de mierda* (ignorant shit), and *necio* (fool) are less frequent but also identified in the corpus.

Table 4.21 shows the statistics of the entities found in OffendES_spans using the SHARE resource. Specifically, 11,035 (33.02% of the corpus) comments contain offensive entities from 33,422 comments in OffendES. In the 11,035 comments, 14,311 non-unique entities

Term	Freq. ↓	Term	Freq. ↓
mierda (shit)	1480	*asco* (disgust)	385
puto (whore)	804	*loca* (crazy)	341
puta (bitch)	706	*gorda* (fat)	336
mala (bad)	510	*coño* (pussy)	331
malo (bad)	442	*basura* (trash)	254
pringada (sucker)	440	*falsa* (false)	239

TABLE 4.20: The 12 most frequent entries of offensive terms in OffendES_spans.

Identification entities/terms	OffendEs_spans
Comments annotated with SHARE	11,035
Unigrams / Uniq. unigrams	13,487 / 636
Bigrams / Uniq. bigrams	582 / 129
Trigrams / Uniq. trigrams	242 / 81

TABLE 4.21: Statistics about entities in the OffendEs_spans corpus using SHARE resource. Uniq: unique (not repeated).

	Comments	Unigrams	Bigrams	Trigrams
NOF	7,293	8,670	329	83
OFF	3,742	4,817	253	159

TABLE 4.22: Total number of non-unique unigrams, bigrams and trigrams labeled with SHARE in NOF and OFF.

(repeated) are recognized, where 13,487 (94.24%) are unigrams, 582 (4.12%) are bigrams and 242 (1.64%) are trigrams.

In addition, Table 4.22 shows the total number of NOF and OFF comments that contain at least one offensive entity in OffendES. In the NOF comments (7,293), we found 8,670 unigrams, 329 bigrams, and 83 trigrams. Regarding the OFF comments (3,742), a total of 4,817 unigrams, 253 bigrams, and 159 trigrams are found. It should be noted that the proportion of comments labeled with at least one offensive entity is much higher in the NOF class (21.82%) than in the OFF class (11.19%) because OffendES is quite imbalanced in the NOF class which includes expletive language.

The OffendES corpus was compiled based on comments from different social networks (Instagram, Twitter, and Youtube). In Table 4.23 we show the number of entities (unigrams, bigrams, and trigrams) found in the comments categorized by the social media platform. We can observe that the largest number of offensive entities are found on Youtube. Specifically, a total of 11,071 unigrams, 448 bigrams, and 143 trigrams are matched. With a considerable decrease, 1,833 unigrams, 114 bigrams, and 93 offensive trigrams are obtained on Instagram. In the last place, Twitter is the social network with the lowest number of offensive words and expressions including 583 unigrams, 20 bigrams, and 6 trigrams. This result is because of an imbalance in the number of

	Unigrams	Bigrams	Trigrams
Instagram	1,833	114	93
Twitter	583	20	6
Youtube	11,071	448	143

TABLE 4.23: Number of non-unique terms labeled in the different social networks.

comments distributed by the social network, 75% of them correspond to Youtube, 18.6% to Instagram, and 6.4% to Twitter.

Finally, as we annotated bigrams and trigrams in the OffendEs corpus, we observed that there are entities that are overlapped (embedded entities). This is considered a challenge for the NLP entity recognition systems. Specifically, we found 589 unigrams which are contained in bigrams. For instance, the entity *puta* (bitch) and *mierda* (shit) are including in the bigram *puta mierda* (fucking piece of shit), or *retrasado* (retarded) into the bigram *retrasado mental* (mentally retarded). A total of 230 unigrams are contained in trigrams such as *violador* (rapist) into *violador de niños* (pedophile) or *puta* (bitch) in *hijo de puta* (son of a bitch) and 26 bigrams are part of trigrams, for instance, *te den* (fuck you) if part of *que te den* (fuck you).

4.6.3 Experiments and results

After the OffendES_spans creation, we aimed to develop a system to automatically detect toxic spans in offensive and non-offensive comments and observe its performance. The toxic span detection task attempted to perform the NER task by assigning each token a label.

We used the pre-trained BERT model to detect all possible offensive entities included in a text. To develop the experiments, we fine-tuned the BERT Transformer by using the BETO model (trained on Spanish texts) "*bert-base-spanish-wwm-cased*" according to the Huggingface library [133]. Optimization was performed using the Adam optimizer [105] with a base learning rate of 1e-5, a batch size of 8, and a maximum sequence of 256.

Table 4.24 shows the results achieved by the model. As we can see, BERT obtained a 91.01% precision, 91.11% recall, and therefore an F_1 score of 91.07%. The results demonstrate the high capability of the transformer-based model in detecting offensive entities by capturing the semantic and syntactic elements of words from a large number of raw text corpora without human intervention. Therefore, we show the utility of SHARE to automatically annotate a corpus with offensive entities and perform the task of automatic offensive span identification.

Model	P (%)	R (%)	F_1 (%)
BERT	91.01	91.11	91.07

TABLE 4.24: Evaluation results on toxic spans detection task. P: Precision, R: Recall.

After performing a result and error model analysis, we found that due to the difficulties of the large Spanish vocabulary, BERT was not able to identify offensive terms such as *desequilibrado* (unbalanced), *chismoso* (gossip), *viejuna* (oldie) and *rata de alcantarilla* (sewer rat). In some cases, BERT could not correctly match the start and end of the entities, e.g., the gold standard included *inútil de mierda* (useless shit) and the system only predicted the term *mierda* (shit). However, we observed that the use of transfer learning systems has been crucial in automatically identifying new offensive terms, saving the manual time involved. As a result, BERT recognized offensive terms such as *pendejasito* (little asshole), *aburrida* (boring) and *pederastas* (pedophiles) not included in SHARE.

4.6.4 Interpretability for offensiveness classification

In order to observe the validity of SHARE as an interpretability tool for offensive language detection in Spanish, we fine-tuned the BERT model on the OffendES_spans corpus and we analyzed a portion of the corpus to compare the attended words with those matched by SHARE.

After fine-tuning the BETO model, we obtained a 93.95% F_1 for the NOF class and a 62.82% F_1 in the OFF class (these experiments are described in Section 4.5: *OffendES - Binary classification with BETO*), showing a great challenge in the classification of offensive comments. Finally, we achieved a macro-average F_1 of 78.39%.

Regarding the explanation analysis, we used the Local Interpretable Model-agnostic Explanations (LIME) [134] to interpret the individual predictions and to evaluate the confidence of the BERT-based system. LIME is a modular and extensible approach to faithfully explaining the predictions of any model in an interpretable way. Among many other options, LIME provides a score for each feature of the comment. In our experiments, we have employed LIME on the fine-tuned BERT model for binary classification of offensive and non-offensive comments.

The comparison of the attended words by BERT using LIME and the words matched with SHARE is shown in Table 4.25. Specifically, seven tweets correctly classified by BERT as offensive are depicted. As can be observed, in most cases, the offensive words identified by BERT match those recognized by our lexicon including *enano* (dwarf),

ID	BERT-LIME	SHARE
818	Das puta pena dalas lo de siempre You're a fucking pity dalas as usual.	puta, das puta pena bitch, you're fucking pitiful
1227	Maldito enano rikillo Damn dwarf rikillo	maldito, enano Damn, dwarf
1545	presa es donde debes estar, pendeja loca. prison is where you belong, you crazy asshole.	pendeja, loca asshole, crazy
1664	No vales nada así te pudras en la cárcel asqueroso de mierda das asco y el MENTIROSO lo eres tú you are worthless hope you rot in jail you filthy piece of shit you are disgusting and you are the liar.	así te pudras, asqueroso, mierda, asqueroso de mierda, asco, das asco, MENTIROSO hope you rot, disgusting, shit, filthy piece of shit, disgusting, you suck, LIER
2071	Eres mala l una psicópata ladrona de perros You are evil and a psychopathic dog thief	mala, psicópata evil, psychopath
3052	TODOS TE ODIAN BRUJA DIFAMADORA, PUDRETE EVERYONE HATES YOU SLANDEROUS WITCH, ROT AWAY	BRUJA, PUDRETE WITCH, ROT AWAY
3173	Tu si que eres patético gañan You really are a pathetic klutz	patético, gañan pathetic, klutz

TABLE 4.25: Interpretability comparison between LIME on BERT (BERT-LIME column) and offensive terms matched by the lexicon (SHARE column). Words highlighted in blue are those identified as possibly offensive. These tweets are annotated as *offensive* and classified as *offensive* by BERT.

pendeja (asshole), *loca* (crazy), *mierda* (shit), *asco* (disgusting), *mala* (evil), *psicópata* (psychopath), *BRUJA* (WITCH), and *patético* (pathetic). Further, there are some instances where SHARE successfully identified offensive terms but BERT failed. For instance, in tweet number 818 the pre-trained language model identifies the word *pena* *pity* but not the insult *puta* (bitch). Similarly, in tweets number 1664 and 3173 SHARE is able to identify the terms MENTIROSO (LIAR), the offensive expressions *así te pudras* (hope you rot), *asqueroso de mierda* (disgusting piece of shit), *das asco* (you suck) and the swearword *gañan* (klutz). Therefore, we believe that SHARE, in addition to being a helpful tool for explainability, could be incorporated into supervised models to aid classification by developing hybrid methods.

4.7 Conclusion

This chapter presents an important milestone addressed in this thesis, the creation of linguistic resources, mainly for Spanish. These resources are oriented toward offensive language research (classification, regression, and NER tasks) as well as emotion classification tasks. Specifically, we created a lexical resource called SHARE, which contains insults and expressions from Spanish speakers, as well as three different datasets. EmoEvent is a multilingual emotion dataset that allows emotion analysis and offensive language research in both English and Spanish. OffendES, the first large-scale dataset for Spanish offensive language research on three different social media platforms (Youtube, Instagram, and Twitter) allows both classification and regression tasks. Finally, OffendES_spans, the first Spanish corpus labeled with entities, in which we rely on the SHARE lexicon to expand the OffendES corpus and allow performing NER tasks. All of these resources are available to the NLP research community and are aimed at fostering research into the development of computational systems to help combat offensive language in social networks, particularly in Spanish. We hope that the NLP community working on Spanish will benefit from these resources in order to advance the state of the art in offensive language research in this language.

Finally, it is worth noting that these generated resources have become pioneers in the research of offensive language in Spanish, and their findings have been published at major scientific conferences in the field of NLP. EmoEvent [135], SHARE [136] and OffendES_spans [136] at Language Resources and Evaluation Conference (LREC), and OffendES [126] at Recent Advances in Natural Language Processing (RANLP).

Chapter 5

Combining linguistic phenomena through a multi-task approach

This chapter describes the main approach proposed in this doctoral thesis to address the offensive language identification task. Given the difficulties of previous NLP approaches in detecting offensive language encountered during the preliminary studies in Chapter 3, we tried to address some of these challenges by proposing a novel method that uses a transfer learning methodology namely MTL. It relies on approaching different linguistic phenomena involved in the expression of offensive language as tasks in order to simultaneously learn common features among them and improve the generalization of the model. It also benefits from the state-of-the-art pre-trained language models based on the Transformer architecture.

In Section 5.1 we define the main linguistic phenomena that could be involved in the expression of offensive language and we consider it important for addressing the offensive language task in text. These phenomena include affective knowledge from emotion and sentiment analysis, the target to which the offense is directed, constructiveness, and rhetorical figures like sarcasm and mockery, among others.

In Section 5.2 we explain the proposed MTL model architecture along with an overview of the literature on previous MTL approaches for offensive language detection. Furthermore, we perform the initial experiments with the proposed approach on two Spanish corpora on offensive language and compare it against a state-of-the-art benchmark model.

5.1 Linguistic phenomena related to offensive language

In the following, we describe the different explicit and implicit linguistic phenomena that we explore in our study to analyze whether and to what extent they contribute to the detection of offensive language.

5.1.1 Emotions

Ekman [137] defined emotions as a process, a particular kind of automatic appraisal influenced by our evolutionary and personal past in which a set of psychological changes and emotional behaviors begins to deal with the situation. In NLP, emotion classification aims to fine-grained automatic classification of texts on the basis of discrete emotional categories like the six basic emotions (*anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise*) of Ekman [137] and the eight primary emotions by Plutchik [138] who added *trust* and *anticipation* and creates a wheel of emotions shown in 5.1 with the purpose of illustrating how emotions are related. The eight basic emotions are organized into four bipolar sets: joy vs. sadness, anger vs. fear, surprise vs. anticipation, trust vs. disgust. The emotions with no color represent emotions that are a combination of two primary emotions. Since emotions are represented by dimensions, the Plutchick model can be considered a dimensional model from a psychological standpoint. In the NLP area, however, is commonly used as a categorical model. An important application within emotion mining is the detection of offensive language since it is inextricably linked to the emotional and psychological state of the speaker/writer [139]. Martins et al. [83] proposed the following new definition of HS in the scope of emotional analysis: "any emotional expression imparting opinions or ideas - bringing a subjective opinion or idea to an external audience - with discriminatory purposes". According to different studies, offensive language semantics contains a strong tendency to negative emotions [140, 141]. Therefore, negative emotions such as *anger*, *disgust*, *fear*, or *sadness* can be conveyed in the form of offensive language. For instance, in the following text "he's a fucking liar, I will no longer trust his words." the author conveys the emotion of anger manifested by the offensive expression "fucking liar". Similarly, positive emotions like *joy* can be an indicator of the absence of offensive language. Our intuition is that most offensive content contains strong negative emotions, which are usually the most direct clues to detect this problem. Therefore, we assume that emotion analysis plays a crucial role in the detection of offensive language and computational models could benefit from this affective knowledge to enhance the detection of this content.

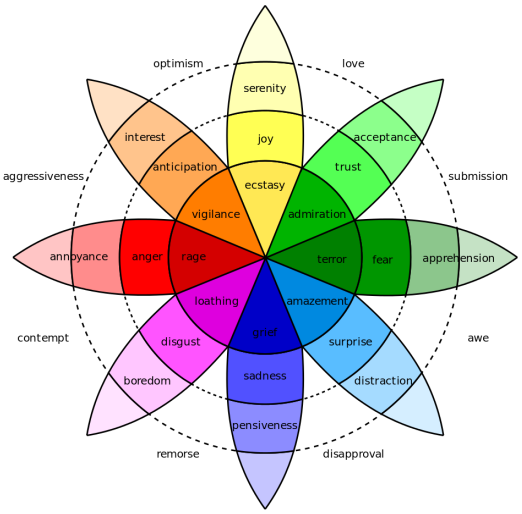

FIGURE 5.1: Plutchik's Wheel of Emotions.

5.1.2 Sentiments

According to the definition in the Cambridge dictionary, sentiment is "a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something". In NLP, sentiment analysis has emerged as one of the most popular areas due to its broad applications in text mining. Its main task is polarity classification and focuses on determining the polarity of a document, sentence, or feature and measuring the degree of the polarity expressed in the document [142]. It is often modeled as a two-class (positive, negative) or three-class (positive, negative, neutral) categorization task. The expression of offensive language and sentiments share some common discourse properties since the former can convey a strong sentiment tendency. In particular, instances with negative sentiments are more likely to involve offensive content. Considering the example "This stupid computer has wiped out all my work for the last week", in addition to expressing anger, it conveys a negative sentiment along with the presence of expletive language targeted at an object ("stupid computer"). Therefore, in this case, the negative sentiment can be an indicator of the presence of offensive language. We assume that the polarity classification task could aid in the identification of offensive language content by both discriminating offensive instances when the opinion is positive and identifying possible offensive instances when the opinion is negative. Note that sentiment analysis is not a "simplified" version of emotion analysis – sentiment analysis is about the expression of an opinion, while emotion analysis is about inferring an emotional private state of a user. These tasks are related, but at least to some degree, complementary [143].

5.1.3 Target

In general, offensive language might be directed at a target (individual or a group of people). Common groups are women, the LGBT community and religious people. The offense target identification is a recent task that has emerged in the last years and can be formulated as a binary/multiclass classification task or as a span identification task. The former focuses on classifying the target of the offenses (individual, group or other) [13] while the latter focuses on identifying the target in the text to which the offense is directed. For example, the following example "The gay community likes to complain and draw attention to themselves" would be classified as offensive to a group and the explicit target of the offense would be "The gay community". As can be seen, the target plays a crucial role in the offensive content and we hypothesize that this knowledge could be a clear indicator to detect the presence of this problem, especially when targets belong to protected classes such as sex, race, age, disability, color, creed, national origin, religion, or genetic information.

5.1.4 Constructiveness

According to Kolhatkar and Taboada [144], constructive comments "intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response. They are typically targeted to specific points and supported by appropriate evidence". These comments can be also referred to as high-quality comments that contribute to the conversation [145]. Kolhatkar and Taboada [144] claim that it is important to consider the constructiveness of comments along with offensiveness when filtering them, as aggressive constructive debate might be a good feature of online discussion. In principle, one might expect there to be a strong negative relationship between constructiveness and offensive content, since constructive comments tend to be non-offensive and vice versa. This would mean that we could rely on existing constructiveness detection systems to filter offensive content. Therefore, our goal is to understand the connection between these two phenomena to help automatic systems to detect offensiveness more accurately. For example, in the following instance "Obviously, but to judge a situation it is important to point out certain things. Context has a lot of weight in these things", the user is giving a constructive response to another user in a conversation, so it can be a clue to the system to probably discard the offensive content.

5.1.5 Figures of speech

We focus on exploring two types of figures of speech that could be involved in offensive content: Mockery and sarcasm. Mockery is the act of insulting or making light of a person usually in a cruel and hurtful way that highlights unflattering characteristics. An example of mockery with the intent to hurt, intimidate or frighten is called bullying, a clear case of offensiveness. For instance, in the text "He is as peaceful as anchovy is mammalian." there is a mockery directed at a person by means of an analogy, thus, this literary device is strongly involved in the expression of offensiveness when the intention is negative. Similarly, sarcasm is a form of irony (implying the opposite) that is directed at a person or group, with the intent to criticize in an offensive or derogatory manner. It aims to attack, being more closely related to a negative mood. Content with sarcasm could be harmful even when all words are polite, and vice versa. For instance, the text "Always such a hard worker!" is apparently not offensive, but depending on the intention and tone, it can be used to criticize the other person's vagueness. Therefore, we consider that it is essential to detect these literary figures in the text that might mask offensive content. Note that the identification of these linguistic phenomena is considered a challenge for the NLP systems and even for humans due to its implicit nature [146].

5.1.6 Profanity language

Many profanity words are used as insults to belittle or offend a person. Among many others, swear words are used in reference to physical, mental, and moral appearance and qualities, personality, sexual orientation and ability, family, racial, national or local origin, religion, beliefs, opinions, and affiliations (political, sports), socioeconomic status, etc. The offensive language used to be expressed with statements that are disrespectful or scornful through insults, expletive language, swear or profanity words/expressions. Thus, profanity language involved in a negative context is an explicit marker of offense [147] and we assume that a detection system for these profanity words/expressions might help to identify offensive content.

5.2 Proposed multi-task learning model

5.2.1 Introduction

As discussed in Section 2, the results obtained by the NLP community so far have shown that offensive language is a complex task, mainly due to the different forms of language

adopted in different cultures, countries, and languages when expressing offense. Previous studies often rely on traditional neural networks or pre-trained language models based on Transformer architecture to obtain semantic features, ignoring other linguistic phenomena that could be involved in the expression of offensive content, which also makes the performance of these automatic systems unsatisfactory in offensive language detection since the model needs to figure out this associations purely from the training data. To overcome the weaknesses of previous work, we build on the intuition that the expression of offensive language could involve other linguistic phenomena (see Section 5.1) that might help to direct this learning process. Offensive language content is potentially related to *sentiments*, *emotions*, the *target*, *figures of speech*, *constructiveness*, and *insults*. First, sentiment analysis is related to offensive language as it typically contains a negative expression or, at least, intended. Second, offensive language contains expressions of anger and might cause fear or other emotions in a target group. Third, the target could be a crucial element of HS, whether mentioned explicitly or not. Similarly, figures of speech like mockery or sarcasm are implicitly mentioned in offensive content and are often used to mask it. Constructiveness comments tend to be non-offensive and can be used as a filter to discard offensive content. Finally, profanity language involved in a negative context is an explicit marker of offensiveness.

Our main hypothesis is whether an offensive language detection system can be improved by exploiting existing resources that are annotated for the linguistic phenomena mentioned and carrying out joint training of a model for offensive language and these aspects. To operationalize this idea as a computational architecture, we propose an MTL approach based on transfer learning to encompass these aspects and detect more accurately the offensive content in textual units. We conduct initial experiments involving two of the linguistic phenomena explained in section 5.1, namely emotions and sentiments, and evaluate the system on two Spanish corpora on offensive language. We compare our proposed approach against a strong Single-Task Learning model (STL), our baseline, based on the state-of-the-art. These first experiments show that our proposal outperforms a strong baseline and, in particular, is able to leverage information from related linguistic phenomena to cope with common errors performed by the strong baseline, which were also observed during the preliminary research addressed in Chapter 3.

5.2.2 Architecture

In this section, we explain the architecture we have followed to develop the MTL approach proposed in this doctoral thesis. First, we give an overview of the definition of transfer learning. Based on this definition, we introduce the STL and MTL settings.

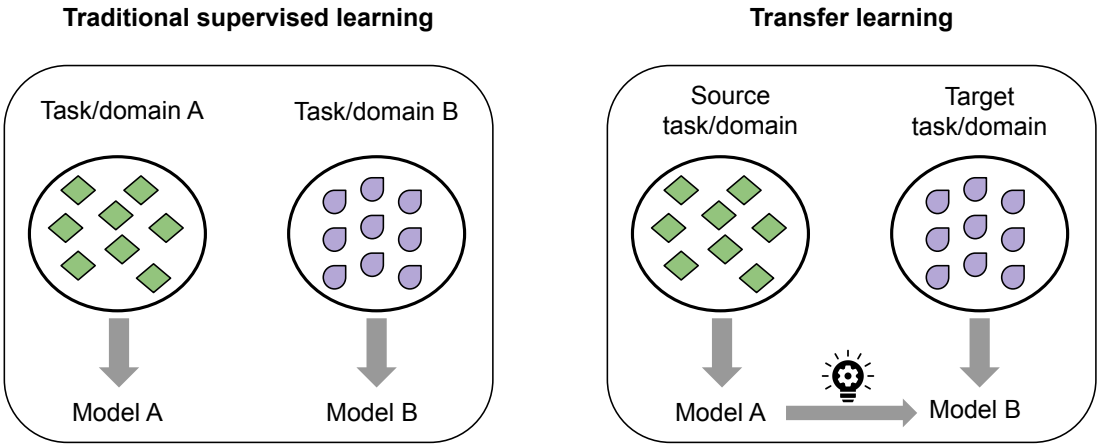

FIGURE 5.2: Traditional supervised learning setup vs Transfer learning setting in ML.

In our methodology, we propose an MTL model to jointly train different linguistic phenomena together with the offensive language task and compare its performance against state-of-the-art STL approaches.

5.2.2.1 Transfer learning

In the traditional supervised learning scenario, a model is trained with labeled data to solve a task in the same domain. Then, we expect this model to work well with unlabeled data for the same domain/task (see Figure 5.2, left side). However, when given data for another task/domain, we require labeled data from the same task/domain again to train a model that performs well on this sort of data. The traditional supervised learning paradigm fails when there is insufficient labeled data for the desired task or domain to train a robust model. Therefore, these types of models lack the ability to generalize to any task beyond the one they learned during the training process. Inspired by how humans are able to transfer knowledge, the NLP community has turned its focus to transfer learning to overcome these problems.

Transfer learning aims to transfer knowledge from a source setting to a target task or domain. Formally, following the notation of Pan and Yang [148] with the binary classification of documents as a running example, transfer learning involves the concepts of a domain and a task (see Figure 5.2, right side). Given a source domain D_S and learning task T_S, a target domain D_T and learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S = D_T$ or $T_S = T_T$.

In this doctoral thesis, we follow the transfer learning taxonomy defined by Ruder [149] to introduce the MTL scenario. In the scope of transfer learning, when there are different tasks and labeled data in a target domain, the taxonomy refers to inductive transfer learning. Within inductive transfer learning, there are two possible scenarios: sequential learning (if the tasks are learned sequentially), and MTL (if the tasks are learned simultaneously). In our study, we focus on the MTL setting to address the offensive language task. Before introducing the MTL scenario, it is important to understand the concept of the STL approach to observe the differences between both setups.

5.2.2.2 Single-task learning

STL is a setting that updates the weight of neural networks using the input sequence of a single classification task in which a labeled dataset is used. In this setting, only a loss function is involved to optimize the task in question. In our methodology, we rely on this setting to establish a baseline and compare its performance with the proposed MTL scenario. To this end, we use an STL model that uses the offensive language task in question as the sole optimization objective. This setting follows a state-of-the-art architecture based on Transformer which is based on an attention mechanism that learns contextual relations between words (or sub-words) in a text. Two separate mechanisms are involved in this architecture: an encoder that reads the text input and a decoder that produces a prediction for the task [1]. Unlike directional models, where the text input is read sequentially, the Transformer encoder reads the whole word sequence at once, allowing the model to learn the context of a word on the basis of its entire surroundings. Our STL setup falls into the transfer learning methodology, as we rely on a state-of-the-art self-supervised language model that operates under a pre-training and fine-tuning paradigm: the model is first pre-trained over a large text corpus and then fine-tuned on a downstream task. Specifically, in the experiments addressed with this setting along this doctoral thesis, we rely on a state-of-the-art pre-trained language model namely Bidirectional Encoder Representations from Transformers (BERT) [39] and fine-tune it on the offensive language task in question.

5.2.2.3 Multi-task learning

While acceptable performance can generally be achieved by focusing on a single task, one might ignore important information that could help to further improve task performance. In particular, this information might arise from the training feedback on related tasks. This is where the MTL paradigm comes in. Within the inductive transfer learning taxonomy, the MTL aims to use the process of learning multiple tasks in order to

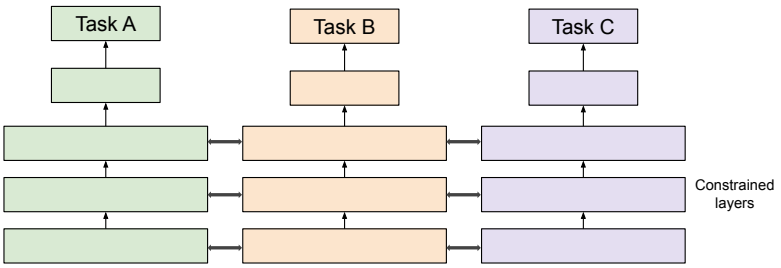

FIGURE 5.3: Soft parameter sharing.

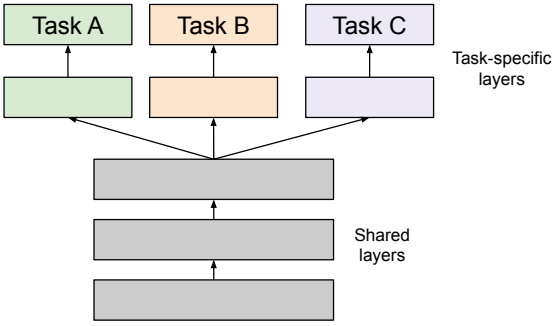

FIGURE 5.4: Hard parameter sharing.

improve the performance on each task [150]. These tasks are usually related and share some commonalities, though they may have different data or features. When the model learns these tasks, some clues from one task can be used to improve the other by sharing features. Therefore, we can enable our model to generalize better on our downstream task.

In the field of DL, MTL is typically implemented with either *hard* or *soft parameter sharing* of hidden layers. In the following, these two methods are introduced:

- **Soft parameter sharing**. In this setting (see Figure 5.3), each task has its own model with its own parameters, and the distance between the model parameters of tasks is added to the joint objective function. The distance between the parameters of the model is regularized in order to encourage similarity between the parameters.
- **Hard parameter sharing**. This technique is the most widely used approach to MTL in neural networks and was introduced by Caruana [150]. It consists of a single encoder that is shared and updated between all tasks, while keeping several task-specific output layers [149] (see Figure 5.4). With this technique, it is possible to reduce the risk of overfitting. This fits when we consider that the more tasks learned concurrently, the more the model has to find a representation that captures all of the shared information, and the lower the likelihood of overfitting in the original task.

A series of mechanisms make the MTL approach plausible for increasing the generalization and performance of deep neural networks. These mechanisms have been identified by Caruana [150] and are described in the following:

- **Implicit data augmentation**. The data used to train the MTL model is going to increase significantly as we need as many labeled datasets as tasks involved in the training process. This fact will help to increase the generalization of the MTL model since learning just one task bears the risk of overfitting to the patterns of that task.
- **Attention focusing**. When the data available for a task is very noisy or limited, it may be a constraint for the model to differentiate between important and not important patterns. MTL may assist the model in focusing its attention on those patterns that are related to the task we want to address in question, as related tasks could bring additional knowledge concerning the relevance or irrelevance of those patterns.
- **Representation bias**. Since tasks with differently labeled data are involved in the process of learning, it will help the model to generalize to new tasks as long as they are related, since the model can learn more robust and universal representations.
- **Regularization**. The MTL model acts as a regularizer. By having only one task in the training process, it increases the possibility of the model to learn noise in the training data and, as soon as the model is exposed to new data, its performance decreases. However, when several tasks are related, the risk of overfitting decreases considerably allowing the model to generalize.

Based on these benefits, we decided to use the MTL scenario in a hard parameter sharing setting to address the offensive language detection task. We propose to train a model concurrently for different tasks that could be involved in the expression of offensive language while using a shared representation. Our goal is to develop a system to enable the model to recognize common features that occur in offensive messages and through these features be able to more accurately identify this phenomenon.

Specifically, the MTL architecture we develop can be observed in Figure 5.5. We fine-tune a Transformer model jointly on related tasks (linguistic phenomena that could be involved in offensiveness) together with the downstream task (offensive language detection). First, the input is tokenized using the WordPiece tokenization algorithm, an extension of byte pair encodings. The sequence has two special segments, the first token [CLS] that contains the special classification embedding and the [SEP] token that is used to separate sentences. The first token is used in the final hidden layers as the

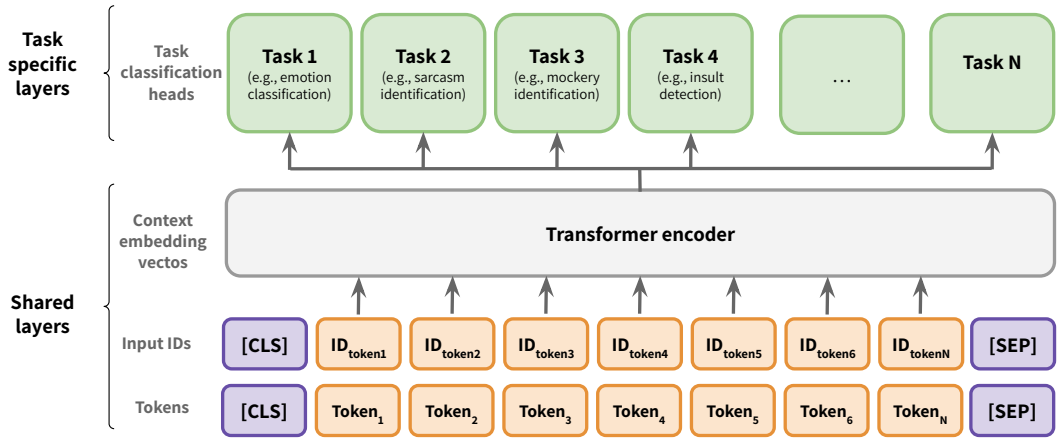

FIGURE 5.5: Proposed MTL to evaluate the impact of including related phenomena to offensive language. The input representation is Transformer-based tokenization and each task corresponds to one classification head. Features can flow from one task to another through the shared encoder that is updated during training via backpropagation.

representation of the whole sequence. Then, the Transformer encoder is shared by the tasks involved. After that, we add as many sequence classification heads as tasks to the encoder and fine-tune the model on the tasks in question. In the training phase, each task is considered with the same importance, therefore, the objective function weights each task equally. All these tasks belong to binary or multiclass classification tasks, depending on the different classes involved, therefore, a standard cross-entropy loss function is calculated between the ground label and the predicted label. Finally, in the inference phase, for each instance in the dataset, different predictions are assigned, one for each task.

It should be noted that in our architecture the task components are customizable, i.e., there will be as many tasks as linguistic phenomena we want to involve for the detection of offensiveness. This allows us to have a flexible and robust architecture to analyze which components help more in the detection of this phenomenon, without having to involve all of them at the same time. Furthermore, we adopted this architecture because, depending on the language, labeled datasets may not be available for a given task related to a given linguistic phenomenon. Similarly, depending on the offensive scenario (sexism identification, HS detection, toxic language detection), we believe that some linguistic phenomena might help more than others in detecting offensive language.

5.2.3 Experiments

As an initial experiment, we are going to focus on two affective phenomena that we consider closely related to offensive language: sentiments and emotions. As sentiment analysis and emotion analysis have been shown to be beneficial for offensive language

detection systems and most studies have used this knowledge within an STL model, our proposed approach is to focus on the MTL paradigm to combine them.

For our experiments, we considered corpora related to polarity classification and emotion classification in order to involve these linguistic phenomena in the detection of offensiveness. Specifically, for the polarity classification task, we chose one of the most popular datasets in Spanish, namely International TASS Corpus (InterTASS). It was released in 2017 [151] with Spanish tweets and updated in 2018 with texts written in three different variants of Spanish from Spain, Costa Rica and Peru [152] and in 2019, with new texts written in two new Spanish variants: Uruguayan and Mexican [153]. The corpus released in 2019 is the one used in these experiments. Each tweet was annotated by at least three annotators considering four classes: Positive (P), Negative (N), Neutral (NEU), and none (NONE). For the emotion classification task, we took advantage of one of the multilingual emotion corpus generated in this doctoral thesis, namely EmoEvent (see Chapter 4: "*Resource generation*"). Specifically, we use the Spanish version which consists of 8,409 tweets labeled in different emotion categories: the six Ekman's basic emotions (anger, fear, sadness, joy, disgust, and surprise) plus the "neutral or other emotions" label. As we aim to analyze if and to what extent these linguistic phenomena help in the detection of offensiveness, we chose two Spanish datasets for offensiveness detection. One is HatEval, provided by the organizers of Task 5 in SemEval 2019 [27] and is related to HS. This dataset contains tweets about two targets: immigrants and women which are labeled as hateful if the text contains HS against these targets, and non-hateful if the text does not contain any signal of HS against them. For a detailed description of the HatEval dataset see Chapter 2: "*Literature review*", Section 2.4: *"Corpora for offensive language detection"*. The other dataset, MEX-A3T, provided by the organizers in IberEval 2018 [91], is related to aggressiveness and contains tweets from South America, specifically from Mexico City. Tweets are composed of an identifier (id), the text of the tweet (text), and the aggressiveness mark, being 0 if the tweet is not aggressive and 1 if the tweet is aggressive.

In order to validate our MTL approach, we decided to compare its performance against to an STL model that acts as our baseline. This model used the offensive language task as the only optimization objective. In particular, we experiment with a well-known model, named BERT [39]. At the time of this study, there were two variants of BERT trained with Spanish texts: the multilingual BERT (mBERT) and BETO. [112]. mBERT was pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages, including Spanish but it does not provide a language detection mechanism therefore the word piece tokenizer could confuse languages. Moreover, it does not have any explicit procedures to encourage translation equivalent pairs to have similar representations. For this reason, we decided to use BETO as our baseline since it is mainly trained on

Spanish data. We refer to this baseline as STL_{BETO}. Specifically, we use the BETO cased checkpoint[1] from the Hugging Face library. In our experiments, we address two STL_{BETO} scenarios to detect offensiveness, one for the HatEval dataset and the other one for the MEX-A3T dataset.

Regarding the experiments for the MTL approach, we used related tasks (polarity classification, and emotion classification) whose datasets shared the same source of data: Twitter. The aim is to check whether the use of polarity and emotion classification tasks assists in the identification of offensiveness. For this scenario, we make use of the same pre-trained model of the STL scenario: BETO.

In summary, we evaluated both HatEval and MEX-A3T datasets conducting four different experiments for each dataset:

1. In order to obtain the baseline, we evaluate the corresponding offensive language dataset with the Transformer-based BETO model, namely STL_{BETO} method.
2. We perform the MTL approach proposed. Specifically, we experimented with three different configurations to detect offensiveness:
   (a) We train concurrently the model on the polarity classification task and offensive classification task and evaluate it on the corresponding offensive dataset. We refer to this model as (MTL_{sent}).
   (b) We train simultaneously the model on the emotion classification task and offensive classification task and evaluate the model on the corresponding offensive. This model is referred to (MTL_{emo}).
   (c) We train jointly the model on the polarity classification task, emotion classification task, and offensive classification task and evaluate it on the corresponding offensive dataset. We refer to this model as ($\text{MTL}_{sent+emo}$).

We use grid search to tune the hyperparameters of the models on the development sets of the offensive language detection datasets (HatEval and MEX-A3T). Across the two STL_{BETO} experiments, MEX-A3T was fine-tuned for three epochs, the learning rate was set to 4e-05, and the batch size to 16. For HatEval, we follow the hyperparameters used in the study explained in Chapter 3 (Plaza-del-Arco et al. [95]): the epochs were set to 3, the batch size to 16, and the learning rate to 2e-05. For the proposed MTL settings, in the case of HatEval, we trained the model for two epochs, the learning rate was set to 4e-05 and the batch size was set to 32. For MEX-A3T, the model was trained for three epochs, the learning rate was set to 3e-05 and the batch size was set to 16.

[1] `https://github.com/dccuchile/beto`

In order to optimize both approaches STL_{BETO} and MTL in both datasets we use the Adam optimizer and the epsilon was set to 1e-8.

5.2.4 Result analysis

In this section, we report the performance of our methodology along with the comparison with the latest state-of-the-art studies. In order to accomplish this, we have employed the usual metrics in NLP classification tasks, including precision, recall, F_1-score, and macro scores of these metrics.

5.2.4.1 Single-task learning vs. Multi-task learning

We compare the performance of our baseline (STL_{BETO}) with the proposed MTL configurations on HatEval and MEX-A3T datasets. The results are reported in Table 5.1 which shows the prediction performances of each model, and each dataset. For the baseline experiments, in the case of HatEval, we use the results obtained in Chapter 3 after applying the Transformer-based model BETO on the HatEval dataset, and in the case of MEX-A3T, we report the results obtained with STL_{BETO}. The performance of the baseline experiments in both datasets is very promising and shows that the STL_{BETO} model works very well when fine-tuned on a small Spanish dataset. Specifically, in the MEX-A3T task, STL_{BETO} achieved high results with a macro F_1-score of 85.51%, compared to the result obtained in the HatEval dataset, 77.62%. In both datasets, we can see that the most challenging class to identify correctly by STL_{BETO} is class 1 (HS and Aggressiveness). This behavior has been observed during the participants' results in workshops related to the HS detection task.

Regarding the results obtained by the different settings of the MTL model proposed, it is worth noting that, for both HatEval and MEX-A3T tasks, all the MTL configurations (MTL_{sent}, MTL_{emo}, and $\text{MTL}_{sent+emo}$) succeeded in surpassing our baseline STL_{BETO} in terms of macro-P, macro-R, and macro-F_1. In particular, for HatEval the best configuration is MTL_{emo} while for MEX-A3T the MTL_{sent} model achieves the best macro-F_1 score. We suppose that as the MEX-A3T dataset contains tweets written in Mexican, it benefits from the dataset used for the sentiment task (InterTASS) which contains texts written in different variants of Spanish including Mexican. Therefore, a deeper knowledge of this linguistic variant is obtained. Concerning $\text{MTL}_{sent+emo}$ model, it behaves in the same way in both datasets, as shown in Table 5.1. Observing the performance of the model for class 1, it is worth mentioning that our proposal $\text{MTL}_{sent+emo}$ outperforms the precision of STL_{BETO}, and achieves a significantly higher recall by increasing 4.09

		Class 0			Class 1			Macro-Avg		
Dataset	**Model**	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**
HatEval	STL_{BETO}	86.16	74.15	79.70	69.28	83.03	75.53	77.72	78.59	77.62
	MTL_{sent}	87.34	73.40	79.77	69.14	84.85	76.19	78.24	79.13	77.98
	MTL_{emo}	87.53	**74.68**	**80.60**	**70.18**	84.85	76.82	78.85	79.76	**78.71**
	$MTL_{sent+emo}$	**88.92**	72.55	79.91	69.03	**87.12**	**77.03**	**78.97**	**79.84**	78.47
MEX-A3T	STL_{BETO}	91.97	91.15	91.56	78.59	80.33	79.45	85.28	85.74	85.51
	MTL_{sent}	93.39	90.93	**92.14**	78.94	84.09	**81.43**	**86.17**	87.51	**86.79**
	MTL_{emo}	92.36	**91.33**	91.84	**79.14**	81.33	80.22	85.75	86.33	86.03
	$MTL_{sent+emo}$	**93.69**	90.21	91.92	77.83	**84.97**	81.25	85.76	**87.59**	86.58

TABLE 5.1: STL_{BETO} and MTL settings results on the Spanish HS datasets. Class 0: non-HS or non-Aggressiveness, Class 1: HS or aggressiveness. Results that outperform the baseline STL_{BETO} model are in bold. P: Precision, R: Recall.

points in the case of HatEval and 4.64 points in MEX-A3T. This observation is remarkable since the $MTL_{sent+emo}$ succeeds at enhancing particularly the most challenging class (1), by detecting the HS and Aggressiveness tweets that STL_{BETO} was not able to identify.

5.2.4.2 Comparison to the state-of-the-art systems

Table 5.2 summarizes the comparative results of previous studies for Spanish HS detection. In particular, we have selected the state-of-the-art systems which have evaluated both HatEval and MEX-A3T datasets.

Regarding HatEval, in Table 5.2 we show the top three teams' [109], [117], [75] that achieved the best results in SemEval-2019 Task 5 as well as other systems that outperformed the results of the competition. The best model in SemEval-2019 Task 5 was presented by [109] where authors obtained a macro-F_1 score of 73.0% using a linear kernel SVM trained on a text representation composed of a bag of words, a bag of characters, and an embedding of tweets computed from fastText sentiment-oriented word vectors. The system proposed by [117] was based on a linear kernel SVM and focused on a combinatorial framework used to search for the best feature configuration among a combination of linguistic pattern features, a lexicon of aggressive words, and different types of n-grams (characters, words, POS tags, aggressive words, word breaks, function words, and punctuation symbols). They obtained a macro-F_1 score of 73.0%. [75] achieved a macro-F_1 score of 72.9%, presenting a pre-trained BERT model on Twitter data and using a corpus of tweets collected over the same period of time from the HatEval training dataset. Another study to consider in SemEval-2019 Task 5 was the system presented by [92] which incorporated sentiment features. This system achieved a macro-F_1 score of 72.5%, implementing a linear SVM model based on linguistic features, semantic similarity with a domain-oriented lexicon, sentiments, word embeddings, topic modeling, and TF-IDF n-grams of words and characters. On the other hand, we found

Dataset	System	Class 0	Class 1	Macro-F_1
HatEval	SVM with sentiment features [92]	75.3	69.8	72.5
	BERT [75]	73.0	72.7	72.9
	SVM with features [117]	76.1	69.9	73.0
	SVM with fastText sentiment embedding [109]	74.9	71.1	73.0
	multi-channel BERT [78]	-	-	76.6
	BETO [95]	79.7	75.5	77.6
	MTL$_{sent+emo}$ (Proposed approach)	79.91	**77.03**	**78.47**
MEX-A3T	EvoMSA7 [154]	89.33	74.68	82.00
	Ensemble BETO models and adversarial data augmentation [155]	**91.95**	79.98	85.96
	BETO [156]	91.07	79.69	85.38
	MTL$_{sent+emo}$ (Proposed approach)	91.92	**81.25**	**86.58**

TABLE 5.2: Comparative results for the HS detection task in Spanish. Results on classes 0 and 1 are in terms of F_1-score.

the study of Sohn and Lee [78]. It used a Transformer-based system, the multilingual BERT model trained for several languages. After analyzing these studies, as can be seen in Table 5.2 it is worth mentioning that our proposed model MTL$_{sent+emo}$ significantly outperforms the best result of the SemEval-2019 Task 5 by 5.47 points in terms of the macro-F_1 score and also slightly outperforms the results of the study performed in Chapter 3 using this dataset. Moreover, it should be noted that our model successfully detected the HS class obtaining an F_1 score of 77.03%. Related to the MEX-A3T dataset, [154] proposed a text classifier that combines two models called B4MSA and EvoDAG. B4MSA is a minimalistic classifier independent from domain and language and EvoDAG is a classifier based on Genetic Programming. [155] and [156] used the BETO model trained specifically for Spanish and similar to the studies of HatEval, it improves also the previous system. Our proposed model MTL$_{sent+emo}$ achieved the best results by obtaining a macro-F_1 score of 86.58%. Similar to the previous dataset, our model successfully identified the HS class obtaining an F_1 score of 81.25%.

5.2.4.3 Knowledge transfer from sentiment and emotion analysis

Our results show that the knowledge from sentiment and emotion classification improves offensive language detection on both HatEval and MEX-A3T datasets. Table 5.3 introduces examples of improvements in HatEval achieved by the MTL$_{sent+emo}$ system, over the STL$_{BETO}$ model. As the affective classification tasks lead the MTL model to learn how to predict the polarity and emotion labels for the instances, the representations computed by the encoder embed the affective knowledge. This allows the MTL$_{sent+emo}$ model to classify HS more accurately by leveraging the affective nature of the instance.

	Tweet	**Gold**	**STL$_{BETO}$**	**MTL$_{sent+emo}$**		
				HS	**Sent.**	**Emot.**
1	*Enseñando a ser puta yo las amo* (Teaching to be a whore I love them)	0	1	0	P	joy
2	*puta la madre, tu eres una mujer muy guapa @user* (fucking hell, you are a very beautiful woman @user)	0	1	0	P	joy
3	*@user Que ganen la sudaca, amén* (@user May they win the sudaca, amen)	0	1	0	P	joy
4	*ANA, tu eres una GUERRERA, no te vamos a dejar abandonar este barco PUTA AMA, te quiero mi niña* (ANA, you are a WARRIOR, we will not let you leave this ship FUCKING BOSS, I love you my girl)	0	1	0	P	joy
5	*Redadas y devoluciones en caliente y fríoo! Estamos hasta los CO-JONES. Más de 100 inmigrantes hieren a siete guardias civiles con ácido y cal viva para saltar la valla de Ceuta* (Hot and cold raids and returns! We are fucking done. More than 100 immigrants hurt seven civil guards with acid and quicklime to jump over the fence in Ceuta)	1	0	1	N	anger
6	*La versión sudaka del Isis. Me parece que cambiemos tendría que abrir un poco más los ojos y tomar más en serio las acciones estos parásitos subordinados* (The sudaka version of Isis. It seems to me that we should change and open our eyes a little more and take the actions of these subordinate parasites more seriously)	1	0	1	N	anger
7	*Te vengo a enseñar y a educar que vos puta a mi no me vas a ganar, también rima?* (I come to teach you and to educate you that you whore to me you are not going to win, also it rhymes?)	1	0	1	N	anger
8	*400 voltios y que quiten las concertinas, y el que tenga huevos que salte* (400 volts and that they remove the razor wire fences, and whoever has balls should jump)	1	0	1	N	anger
9	*Por desgracia, no queda otra, aportan poco y nos cuestan mucho, incluido nuestra seguridad* (Unfortunately, there is no other way, they contribute little and cost us a lot, including our security)	1	0	1	N	sadness

TABLE 5.3: STL$_{BETO}$ vs. MTL$_{sent+emo}$ samples from HatEval dataset, showing improved MTL performance. P: Positive, N: Negative. English translation of Spanish tweets is provided between brackets.

Looking at the examples in Table 5.3 it is important to point out that people often use some expressions that contain offensive words, however, the expression is not necessarily offensive since it conveys a positive polarity and emotion. For instance, tweet number 4 contains the expression *puta ama* (fucking boss) which is positive although the presence of the offensive word *puta* (whore) is used. In this case, the STL$_{BETO}$ model mislabeled the tweet as HS, whereas the MTL classified it as non-HS since the polarity and emotion predicted were *positive* and *joy*, respectively. Similarly, tweets 1, 2, and 3 with *positive* polarity and *joy* emotion were correctly classified by our proposed model as non-HS but not by the STL$_{BETO}$ model which prediction was HS. These examples, as well as the expressions, also contain offensive words associated with misogyny and xenophobia, but the emotion they evoke is *positive*. The rest of the examples with *negative* polarity and conveying *anger* and *sad* emotions were misclassified by STL$_{BETO}$ but not by the MTL$_{sent+emo}$ model. For instance, in tweet 5 with the negative words *redadas* (raids) and *hieren* (hurt), tweet 6 with *parasitos subordinados* (subordinate parasites), and tweet 8 with the expression *el que tenga huevos* (whoever has balls), it is again shown that MTL benefits from the affective knowledge learned from sentiment and emotion

Class	STL_{BETO}		$\text{MTL}_{sent+emo}$	
	non-HS	HS	non-HS	HS
non-HS	685	255	682	258
HS	104	556	85	575

TABLE 5.4: Confusion matrix of HatEval.

classification tasks.

5.2.5 Error analysis

In order to gain deeper insight into the proposed MTL model performance, we conducted an error analysis from both quantitative and qualitative levels. We mainly analyzed the instances in the test set that were wrongly labeled by the STL_{BETO} and the $\text{MTL}_{sent+emo}$ models in HatEval dataset. Since the gold labels of the MEX-A3T test set are not publicly available, we have not performed this analysis for this dataset.

Based on quantitative analysis, we analyzed the confusion matrices of STL_{BETO} and $\text{MTL}_{sent+emo}$ models and compare them in Table 5.4. The MTL system in HatEval mislabeled only 85 HS instances to Non-HS compared to 104 instances misclassified with the STL_{BETO} model. As we highlight in the analysis of the results, it shows the MTL model's ability to distinguish the hateful text by reducing the number of false positives in HS class. On the other hand, the $\text{MTL}_{sent+emo}$ system mislabeled 258 non-HS instances to HS compared to 255 instances misclassified with the STL_{BETO} model. Therefore, we consider that the knowledge provided by the external datasets (InterTASS and EmoEvent), although very slightly detrimental to the Non-HS class, not just improves the prediction in general, but the performance in the HS class is particularly enhanced.

Concerning the qualitative analysis, we focused on analyzing some tweets misclassified by the $\text{MTL}_{sent+emo}$ system to identify the possible challenges that the system faces with the Spanish and HS detection. In addition, we also looked at the predictions related to emotion and polarity classification in order to analyze how they contribute to the detection of HS. We select some mislabeled tweets predicted by the $\text{MTL}_{sent+emo}$ system: two false positive and two false-negative tweets which can be seen in Table 5.5. In the first false positive, there is the presence of the xenophobic word *negrata* (nigga) but at the same time, the user includes the positive word *top* (top). Consequently, due to the lack of context and the short length of the tweet, the system is labeled as HS. Moreover, it should be noted that this tweet is classified as *neutral* polarity and *others* emotion which shows that is not clear if the tweet is positive or negative and therefore it

Tweet	**Gold label**	**MTL$_{sent+emo}$**		
		HS	**polarity**	**emotion**
Para mi con el negrata la delantera es top (For me with the nigga in the front is top)	0	1	NEU	others
O sea, ¿se supone que no puedo tener opiniones? ¿debo estar siempre callada porque soy una mujer? me dices puta porque expreso lo que pienso, supongo que prefieres que me limite a sentarme y sonreír. Cuando una mujer se harta y contraataca de repente el macho no sabe como actuar. (I mean, am I not supposed to have opinions? should I always be quiet because I am a woman? you call me a whore because I express what I think, I guess you prefer me to just sit and smile. When a woman gets fed up suddenly man doesn't know how to act)	0	1	N	anger
Devolución exprés ahora y siempre. Y más concertinas y lo que haga falta para que no entren (Express return now and always. And more razor wire fences and whatever else is needed to keep them out)	1	0	N	others
Quitar las concertinas y poner ametralladoras (Remove the razor wire fences and set up the machine guns)	1	0	N	others

TABLE 5.5: Tweets mislabeled by the MTL$_{sent+emo}$ model. Two false positives and two false negatives, respectively. English translation of Spanish tweets is provided between brackets. N: Negative. NEU: Neutral.

demonstrates the difficulty of detecting the HS by the MTL$_{sent+emo}$ model. In the second false positive, the user employs the misogynist word *puta* (whore), but without offending anyone, because she is referring to herself and is annoyed with the judgments that other people make about women. It is possible that the MTL$_{sent+emo}$ model mislabeled it because there is a misogynist word and also the emotions and sentiments expressed in the tweet are negative, two factors that, together, are usually associated with the presence of HS. In the case of the false-negative instances, the system has may predict it wrongly because although the user is expressing xenophobia, there is no explicit mention of immigrants in the tweet, which is a challenge for NLP systems due to the implicit information. In addition, the polarity is classified correctly but the emotion predicted is other which shows the difficulty of detecting HS in this tweet.

Finally, it is worth noting that we have detected some mislabeled tweets in the datasets, which complicates the learning process of the models.

5.3 Conclusion

In this chapter, we introduced the main approach presented in this doctoral thesis for the task of offensive language detection. Our study builds on the assumption that the discourse of offensive language could involve other linguistic phenomena, and might be directed toward a specific individual or group. In order to operationalize this idea

as a computational architecture, we decided to develop an MTL scenario to explore if training a model simultaneously for different tasks related to offensive language detection is helpful for the purpose of offensive language detection. In this scenario, different NLP tasks (linguistic phenomena that could be involved in the expression of offensiveness) are jointly trained together with the downstream task (offensive language detection), which allows the model to better generalize on this downstream task. To validate our hypothesis, we conduct the first experiments on two Spanish corpora related to offensive language: HatEval and MEX-A3T. In these experiments, we incorporate two affective phenomena that can help to address the offensive language detection task: sentiments and emotions and we have used corpora labeled for each of the tasks to include them in the MTL setting. Moreover, we have explored which combination of these phenomena could be the most successful. Experiments conducted on two benchmark corpora show the efficacy of our proposed approach in achieving convincing performance over an STL baseline based on the state-of-the-art. The performance achieved by our proposed model and a detailed knowledge transfer analysis from sentiment and emotion analysis shows that polarity and emotion classification tasks help the MTL model to classify offensive language more accurately by leveraging affective knowledge. In particular, we found that the model is good at retrieving HS tweets not identified by BERT. A plausible scenario here is that negative sentiments and negative emotions are associated with the general spirit of offensive language so the presence of these indicators permits the model to predict the presence of offensive language more accurately.

In the next chapter, we will evaluate the MTL approach in different offensive scenarios (sexism identification, toxic detection, and HS detection) involving not only the affective knowledge from sentiments and emotions that we have employed in the first experiments of this chapter, but also the rest of the related phenomena described in Section 5.1 such as mockery and sarcasm, constructiveness, profanity, and target.

Chapter 6

Detection of offensive scenarios using the multi-task approach

In this chapter, the main approach proposed in the thesis, namely MTL that combines different linguistic phenomena (see Chapter 5: *"Combining linguistic phenomena through a multi-task approach"*) is going to be used to address different real-word scenarios that involve offensive speech. These scenarios have been proposed in recent NLP evaluation campaigns and include the detection of toxicity in Spanish comments (DETOXIS) [19], the hate speech and offensive contents identification (HASOC) [33] and the sexism identification in social networks (EXIST) [44]. For each of the scenarios, we follow a methodology to connect different tasks to the MTL system that represent the linguistic phenomena we want to involve while addressing the specific scenario. These phenomena include affective knowledge from sentiment and emotion analysis; rhetorical figures such as sarcasm, mockery, and irony; explicit expressions like insults or improper language, among others. During the evaluation, we aim to identify which linguistic phenomena help to address the task proposed in each scenario, then, through an extensive analysis we explore how the knowledge is transferred between tasks. Furthermore, we conduct an error analysis examining the challenges faced by the MTL approach proposed.

6.1 Detection of toxicity in comments in Spanish

6.1.1 Problem definition

The task of toxic comment classification in Spanish was proposed for the first time in the DETOXIS (DEtection of TOxicity in comments In Spanish) shared task [19] at IberLEF 2021. The organizers defined a comment as toxic when "it denigrates, hates or vilifies,

attacks, threatens, insults, offends or disqualifies a person or group of people based on characteristics such as race, ethnicity, nationality, political ideology, religion, gender and sexual orientation, among others". This shared task proposed two classification subtasks:

- Subtask 1: A binary classification task that consists of classifying a comment as *toxic* or *non-toxic*.
- Subtask 2: A multiclass classification task aimed to identify the toxicity level of a comment.

The detection of toxicity is a challenging task because the expression of toxic language can be formulated in multiple ways: explicitly (insults, improper language, mockery) or implicitly (sarcasm, irony). In addition, toxic comments may present different levels of intensity in toxicity (from rude and harmful comments to more aggressive ones). However, it is important to invest efforts in the development of automatic NLP-based systems to combat this problem on the web and even more in Spanish, where the number of resources is limited.

6.1.2 Metholodogy

We focused on addressing subtask 1 proposed in the DETOXIS shared task which consists of classifying a comment as *toxic* or *non-toxic*. A first approximation of the MTL methodology was presented in the DETOXIS shared task [157] achieving first place among the participants in both subtasks. However, in this section, we aimed to study in depth this approximation with the objective of analyzing in detail which linguistic phenomena contribute the most to the detection of toxic language. The methodology follow to address this task is described in detail in Chapter 5. In particular, we assumed that considering different linguistic phenomena that could be involved in the expression of toxicity can help automatic methods to detect this type of content more accurately. Therefore, in the MTL model, we fine-tuned a Transformer model jointly on the linguistic phenomena labeled in the NewsCom-TOX dataset including the toxicity label. First, the input is represented by the BETO encoder [112]. Then, we added as many sequence classification heads as tasks to the encoder and fine-tune the model on the tasks in question (binary/multiclass classification tasks). In the prediction phase, for each post in the dataset, different predictions are assigned, one for each task involved.

In order to validate our hypothesis, we compared our MTL proposed system with an empirical upper bound, which is BETO [112], the pre-trained language model on Spanish texts based on the Transformer architecture.

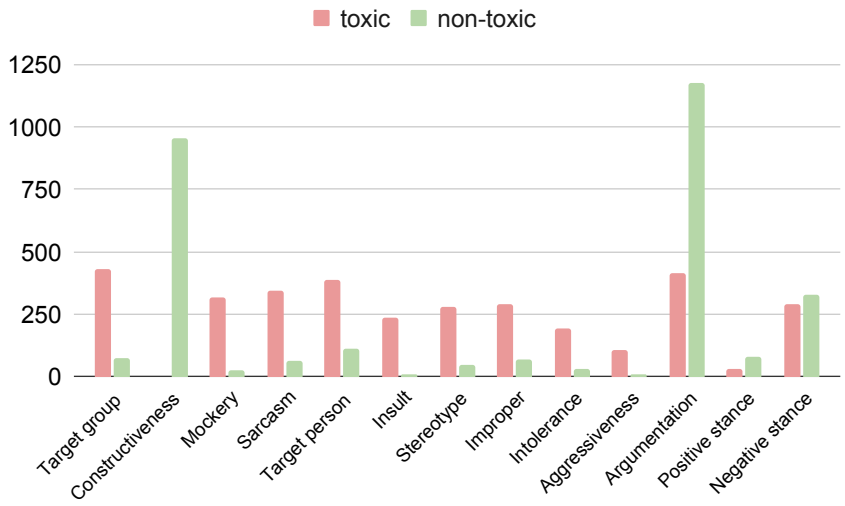

FIGURE 6.1: Distribution of comments by linguistic phenomena in the Spanish NewsCom-TOX training set.

6.1.3 Experimental procedure

In our experiments, we aimed to answer three different research questions: **(RQ1)** which corpora to use to train the different linguistic phenomena? **(RQ2)** which phenomena to include and to what extent do they contribute to the detection of toxic comments?, and **(RQ3)** do these phenomena combined with emotions aid in toxicity detection?

Datasets. To answer the RQ1 question, we selected the NewsCom-TOX corpus annotated with 13 different linguistic phenomena including *argumentation*, *constructiveness*, *positive stance*, *negative stance*, *target*, *stereotype*, *sarcasm*, *mockery*, *insult*, *improper language*, *aggressiveness* and *intolerance* provided by the organizers of the DETOXIS. For a detailed description of the NewsCom-TOX dataset see Chapter 2: "*Literature review*", Section 2.4: *"Corpora for offensive language detection"*. Figures 6.1 and 6.2 shows the distribution of *toxic* and *non-toxic* comments in both training and test subsets partitions, respectively in addition to the rest of the phenomena.

To answer RQ2, we select the linguistic phenomena labeled in the NewsCom-TOX corpus and perform a feature selection method to get an insight into to what extent they contribute to the offensive language detection task. This analysis is performed in Section 6.1.3.1.

To answer RQ3, we chose two recent emotion datasets focus on Spanish texts, among other languages, namely EmoEvent [135], the one we created as a result for this doctoral thesis and Universal Joy [158], focused on Facebook posts labeled with five basic emotions (*anger*, *anticipation*, *fear*, *joy*, and *sadness*). The dataset is a substantially

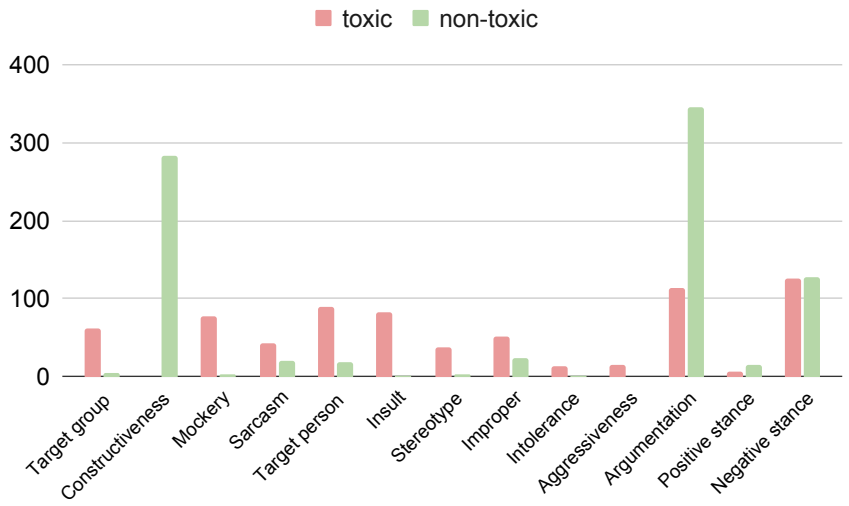

FIGURE 6.2: Distribution of comments by linguistic phenomena in the Spanish NewsCom-TOX test set.

reorganized and cleaned version of one previously described [159] that was collected in October 2014 by searching for public Facebook posts with a Facebook feelings tag.

Data preprocessing. As the NewsCom-TOX corpus contains colloquial comments in response to newspaper articles, we decided to perform different preprocessing steps to reduce the noise of the dataset while training a model. Specifically, we removed URLs and special characters, replaced multiple spaces with a single space, deleted writings with only numbers, and reduced words with more than four repeated characters to three repetitions. For the emotion datasets (EmoEvent and Universal Joy), since the source of comments are social networks, they present numerous challenges in their tokenization, such as user mentions, hashtags, emojis, and misspellings, among others. To tackle these challenges, we normalized all mentions of URLs, emails, percentages, users' mentions, time and date expressions, monetary amounts, and phone numbers. For instance, the token "@user" is replaced by "user". We further normalized hashtags and split them into their constituent words. For example, "#FelizLunes" ("#HappyMonday") is replaced with "Feliz Lunes" (Happy Monday). Finally, emojis were replaced by their aliases using the emoji Python library[1].

Training procedure and hyperparameters. For both baseline BETO and MTL methods, we fine-tuned the models on the combination of the training and development sets provided by the organizers of the DETOXIS shared task for 3 epochs, with a learning rate of $2 \cdot 10^{-5}$ and the batch size of 16. Afterward, the evaluation is carried out on the test set.

[1] `https://bit.ly/3L5yoEd`

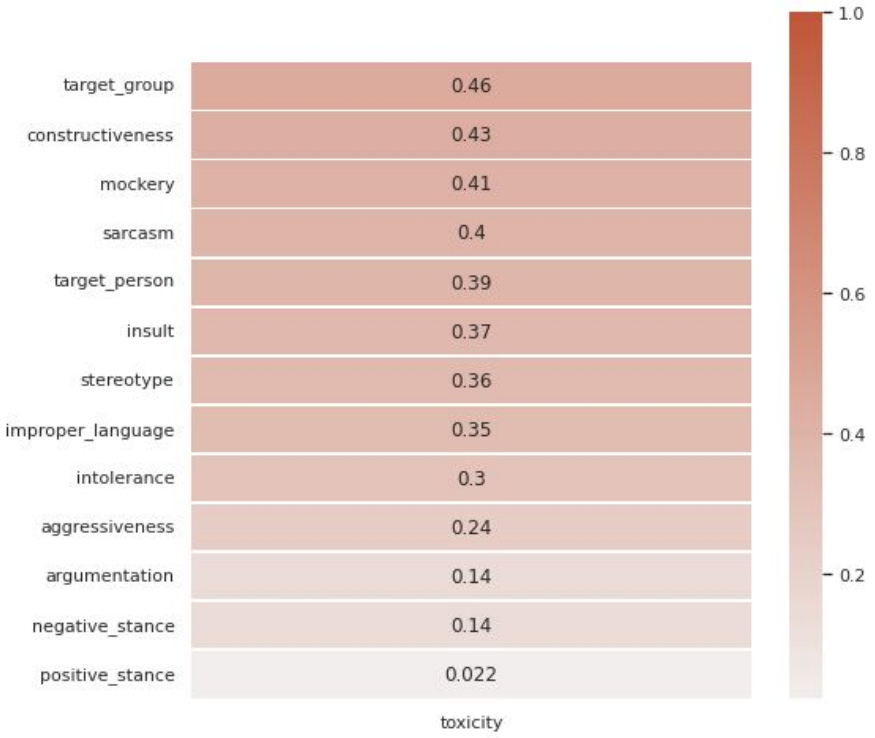

FIGURE 6.3: Results of the mutual information calculation on the linguistic phenomena in the NewsCom-TOX dataset. The coefficient of correlation ranges from 0 to 1: 0 indicates no correlation between the phenomenon and the toxicity class while 1 indicates the opposite.

6.1.3.1 Experiment 1: Analyzing linguistic phenomena related to toxicity

For answering RQ2, we analyzed which implicit and explicit linguistic phenomena are most related to the detection of toxic language. For this objective, we used the feature selection method named mutual information which identifies the most and least relevant features for the classification task (toxic detection). Mutual information is a measure of the mutual dependence between two random variables. The function is based on non-parametric methods for entropy estimation from k-nearest distances [160]. We used this method to analyze the relationship between each of the linguistic phenomena (*target group*, *constructiveness*, *mockery*, *sarcasm*, *target person*, *insult*, *stereotype*, *improper language*, *intolerance*, *aggressiveness*, *argumentation*, *positive stance*, *negative stance*) and *toxicity*. The result of this analysis can be seen in Figure 6.3 where we observed that most of the features are related to the toxicity phenomenon except *the positive stance* whose influence is minimal. In particular, the top 6 most related linked to the toxicity class are target (*target group*, *target person*), *constructiveness*, figurative language (*mockery*, *sarcasm*), and *insult*. However, phenomena such as *argumentation* or *positive stance* are very slightly linked to toxicity.

		Toxic			$\overline{\text{Toxic}}$			Macro-Average		
Model		P	R	F_1	P	R	F_1	P	R	F_1
Baseline	BETO	0.613	0.636	0.624	0.865	0.853	0.859	0.739	0.744	0.742
MTL	MTL_1	**0.620**	0.607	0.613	0.857	**0.864**	**0.860**	0.738	0.735	0.737
	MTL_2	0.608	0.611	0.610	0.857	**0.856**	0.857	0.733	0.733	0.733
	MTL_3	**0.619**	**0.686**	**0.651**	**0.880**	.845	**0.862**	**0.750**	**0.766**	**0.757**
	MTL_4	0.597	**0.670**	**0.631**	**0.873**	0.834	0.853	0.735	**0.752**	0.742
	MTL_5	**0.618**	**0.649**	**0.633**	**0.869**	0.853	**0.861**	**0.743**	**0.751**	**0.747**
	MTL_6	0.607	0.615	0.611	0.858	**0.854**	0.856	0.733	0.735	0.734

TABLE 6.1: Results obtained by incorporating different phenomena as tasks evaluating the MTL model on the NewsCom-TOX test set. Results that outperform the baseline model are in bold. P: Precision, R: Recall.

Label	**Target_gr.**	**Const.**	**Mock.**	**Sarc.**	**Target_per.**	**Insult**	**Ster.**	**Imp. lang.**	**Intoler.**	**Aggr.**	**Arg.**	**Neg.**
MTL_1	✓	✓										
MTL_2	✓	✓	✓	✓								
MTL_3	✓	✓	✓	✓	✓	✓						
MTL_4	✓	✓	✓	✓	✓	✓	✓	✓				
MTL_5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
MTL_6	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

TABLE 6.2: MTL experiments using linguistic phenomena related to toxic language. Target_gr.: Target group, Const.: Constructiveness, Mock.: Mockery, Sarc.: Sarcasm, Target_per.: Target person, Ster.: Stereotype, Imp. lang.: Improper Language, Intoler.: Intolerance, Aggr.: Aggressiveness, Arg.: Argumentation, Neg.: Negative Stance.

After gaining insight into how different linguistic pragmatic strategies are related to *toxicity* in the NewsCom-TOX dataset, we decided to conduct an incremental approach that consists of incorporating the linguistic phenomena as tasks in pairs following the ranking given by the mutual information method. We aimed to analyze to what extent these phenomena aid in the detection of toxic language using our MTL proposed model.

6.1.3.2 Results

Table 6.1 shows the results of the MTL models including the different linguistic phenomena on the NewsCom-TOX test data, both with the macro-average evaluation and the class-specific values in terms of precision, recall, and F_1 scores. These results are on the main task of toxicity detection, but vary the subsets of tasks related to the linguistic phenomena according to Table 6.2. As can be seen, not all MTL models outperform the baseline, which means that not all linguistic phenomena contribute in the same way to the detection of toxicity. The best performance in terms of Macro-F_1 is obtained by MTL_3 which is trained on the top six linguistic phenomena shown in Figure 6.3 (*target group*, *constructiveness*, *mockery*, *sarcasm*, *target person*, and *insult*), followed by MTL_5 trained on all the features except *argumentation* and *negative stance* concepts. It is noteworthy that this model surpasses the baseline BETO by at least 3 percentage points in terms of F_1 for the *toxic* class and the recall for the toxic class is particularly

increased by 5 points. This achievement is important both for practical applications where detecting toxicity is more relevant than detecting non-toxic texts and from a dataset viewpoint, as most resources have a significantly lower label count of *toxic* class. The models MTL_4 and MTL_5 are able of surpassing the baseline system in terms of F_1 of the *toxic* class.

Finally, in order to answer RQ2, our experiments show that the complementarity of different linguistic phenomena helps in the detection of toxicity in the NewsCom-TOX dataset. Specifically, these components are those that could be involved in the expression of the target (target person, target group), figurative language (mockery, sarcasm), and explicit insults which correspond to the six concepts suggested by the mutual information method. It is worth noting that these phenomena are especially helpful in identifying comments from the *toxic* class, which is the more challenging in this type of problem.

6.1.3.3 Experiment 2: Incorporating emotions

In order to answer RQ3, we conducted MTL experiments introducing emotion knowledge as an additional task by leveraging the two Spanish emotion datasets mentioned above. The emotions considered are *anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise*, *anticipation* and *others*.

Our first attempt to analyze whether emotions contribute to the detection of this type of content is incorporating emotion classification as an additional task to the best MTL observed in Experiment 1, which is MTL_3. Additionally, we also experimented with different combinations of phenomena along with emotions to find the best possible combination for the detection of toxic language along with emotions.

6.1.3.4 Results

Table 6.3 shows the results of different MTL models considering the linguistic phenomena previously described along with emotions. These results are on the main task of toxicity detection, but vary the subsets of tasks related. We compared these results with the baseline BETO and with MTL_3 to observe if the emotional knowledge shared across tasks in the MTL approach further improves the best results achieved so far. As it can be observed, adding the emotion classification task to the MTL_3 experiment does not improve the results. Therefore, our second attempt was to incrementally add phenomena as tasks following the same methodology used in Experiment 1. After performing different combinations, we observed that including *improper language* and omitting *stereotype* concept (MTL_3emo_imp) the results further outperform the MTL_3 method in terms

Model		Toxic			$\overline{\text{Toxic}}$			Macro-Average		
		P	R	F_1	P	R	F_1	P	R	F_1
	MTL_3	0.619	0.686	0.651	0.880	0.845	0.862	0.750	0.766	0.757
MTL_3	emo	0.608	0.611	0.610	0.857	**0.856**	0.857	0.733	0.733	0.733
	emo_imp.	**0.665**	0.657	**0.661**	0.875	**0.879**	**0.877**	**0.770**	**0.768**	**0.769**
MTL_4	emo	0.616	0.644	0.630	0.867	**0.853**	0.860	0.742	0.749	0.745
MTL_5	emo	0.602	0.653	0.627	0.869	0.842	0.855	0.736	0.747	0.741

TABLE 6.3: MTL results on NewsCom-TOX test set by incorporating different phenomena as tasks along with emotions. Results that outperform the MTL_3 model are in bold. P: Precision, R: Recall.

of macro-F_1 (0.757 - 0.769) and F_1 in the *toxic* (0.651 - 0.661) and *non-toxic* (0.862 - 0.877) classes. In particular, this model yields fewer false positives in the *toxic* class than the MTL_3 as we can see in the differences in precision scores (0.619 - 0.665). In addition, we also added the emotion classification task to the MTL_4 and MTL_5 but the results do not improve.

Finally, in order to answer RQ3, we concluded that the *target group*, *constructiveness*, *mockery*, *sarcasm*, *target person*, *insult*, and *improper language* phenomena complement the emotion knowledge to help in the detection of toxic language.

Comparison with the DETOXIS 2021 participants' systems

As indicated in Section 6.1.2, we presented at the DETOXIS shared task a first approach of the MTL technique. This preliminary approach took into account the various linguistic phenomena discussed, but without focusing on which of these phenomena contributed the most to the detection of toxic language. As shown in Table 6.4 our participation (SINAI) ranked first among all the participants in the DETOXIS shared task by achieving an F_1-score of 0.6610 in the *toxic* class. This first approach combines all the concepts (constructiveness, argumentation, mockery, sarcasm, positive stance, negative stance, target person, target group, stereotype, insult, improper language, aggressiveness, intolerance, and emotions). Teams GuillemGSubies [161] and AI-UPV [162] ranked second and third, respectively, in the shared task. GuillemGsubies fine-tuned the BETO model using grid search and data augmentation with masked language model substitution. AI-UPV employed classic ML model and Transformer-based models to address the task, obtaining the best result with the BETO model. After our initial approach that achieved the best result in the shared task (SINAI), we aimed to investigate in depth which of the phenomena contributed the most to the toxic language. As a result of our research, we discovered that the best-performed model described throughout this section (MTL3 emo imp) outperformed our first place in the task, achieving state-of-the-art results in the detection of Spanish toxic language. Specifically, we surpassed this initial approach by 1.49% of F_1-score in the *toxic* class. As a result, we believe that an in-depth

study of what features in the MTL approach contribute the most to the detection of toxic language is important in order to overcome this task and that the MTL model that takes these features into account is a successful system that provides an improvement over previous state-of-the-art methodologies for the Spanish language.

Ranking	Team	P	R	F_1
MTL3_emo_imp	Our approach	0.6650	0.6570	**0.6610**
1	SINAI	0.6569	0.6356	0.6461
2	GuillemGSubies	0.7029	0.5234	0.6000
3	AI-UPV	0.6360	0.5672	0.5996

TABLE 6.4: Comparison of our best model (MTL3_emo_imp) with the three best approaches used by the participants in DETOXIS 2021 shared task. Precision (P), Recall (R) and F_1-score in the *toxic* class are reported.

6.1.4 Knowledge transfer from linguistic phenomena analysis

In order to get a better understanding of how the knowledge is transferred across the different tasks in the MTL models for the detection of toxic language, two different model analyses are conducted. In the former, in order to observe how the linguistic phenomena contribute to the detection of toxic language, we show a comparison between some of the examples in which the baseline fails in the prediction but the MTL_3 model correctly identifies the class. In the latter, we compared some examples incorrectly predicted by the baseline BETO but correctly predicted by the MTL_3emo_imp to observe how emotions help in the knowledge shared between tasks.

The first comparison (BETO vs MTL_3) is shown in Table 6.5. Specifically, we show 6 examples, namely 3 false negatives and 3 false positives performed by the baseline BETO model. Regarding the false negatives, in the post with ID 10_302, the MTL system has correctly identified *mockery* ("Tu la cabeza para peinarte" - Your head to comb your hair) - this expression is used to make fun about the intelligence of a person - and is directed to a person, therefore as these aspects are closely linked to the presence of toxic language. We assume that the MTL takes advantage of these aspects to correctly classify the instance. The next sample, a post with ID 14_288, is directed to a person, and contains *mockery* and *sarcasm* ("Tú entiende lo que quieras campeón" - You understand what you want champion) - champion is used sarcastically - correctly identified by the MTL. Instance with ID 18_018 is also correctly predicted by the MTL model as *toxic*. In this case, the post is directed to a group and we believe that the model has benefited from this knowledge to correctly predict the class. Regarding the false positives (posts with ID 14_265, 18_184, 14_287) from BETO, we observed that the phenomenon that has contributed to correctly identifying them as *non-toxic* by the

				MTL_3	
ID	Text	Gold	BETO	Toxic	Features
10_302	*Si no tienen permiso de trabajo ¿donde han estado trabajando? Tu la cabeza para peinarte* (If they don't have a work permit where have they been working? Your head to comb your hair)	Toxic	Toxic	Toxic	target person, mockery
14_288	*creo que te está intentando hacer ver que esos "manifestantes" tenían de pacíficos lo que tienen las anchoas de mamíferos. Pero eh, que tampoco te quiero dar las cosas hechas. Tú entiende lo que quieras campeón* I think he is trying to make you see that those "demonstrators" were as peaceful as anchovies are mammals. But hey, I don't want to give you the facts either. You understand what you want champion	Toxic	Toxic	Toxic	target person, mockery, sarcasm
18_018	*Cuando los españoles no quieren currar pues sí. PER, o cómo vivir del campo sin trabajar* (When Spanish people do not want to work, then yes. PER, or how to live off the land without working)	Toxic	Toxic	Toxic	target group
14_265	*Obviamente, pero de cara a juzgar una situación es importante remarcar ciertas cosas. El contexto tiene mucho peso en estas cosas. En España un chaval de 17 años con un fusil de asalto en una mani es un puto loco. En EEUU los ves por la televisión día si día también, y nadie se lleva las manos a la cabeza ni la policia hace nada en concreto sobre eso* (Obviously, but in order to judge a situation it is important to highlight certain things. The context has a lot of weight in these things. In Spain a 17 year old kid with an assault rifle in a demonstration is a fucking madman. In the USA you see them on TV day after day, and no one raises their hands to their heads, nor do the police do anything concrete about it.)	Toxic	Toxic	Toxic	constructiveness
18_184	*Vas a ir tú al jefe de la plantación a decirle que no contrate negros para contratar a españoles pagándoles el doble? Porque que yo sepa el que decide hacer eso es el empresario, no es culpa de los negros que solo sean ellos los que van a esos sitios a trabajar levantándose a las 4 de la mañana hasta las 8 de la tarde a +40 grados bajo pleno sol en verano* (Are you going to go to the boss of the plantation and tell him not to hire blacks to hire Spanish people and pay them twice as much? Because as far as I know the one who decides to do that is the employer, it is not the fault of the blacks that they are the only ones who go to those places to work getting up at 4 in the morning until 8 in the evening at +40 degrees under the sun in summer.)	Toxic	Toxic	Toxic	constructiveness
14_287	*el chaval oí en un video que decía "i work at that business", me da a entender que uno de los negocios de la zona era donde trabajaba. Y no sé tú, pero la idea de quedarte en el paro en plena pandemia, porque unos sinvergüenzas van a saquear tu empresa... Igual te hace dar ganas de defenderla* (I heard in a video that the guy said "i work at that business", it seems to me that one of the businesses in the area was where he worked. And I don't know about you, but the idea of being unemployed in the middle of a pandemic, because some scoundrels are going to loot your company.... It might make you want to defend it)	Toxic	Toxic	Toxic	constructiveness

TABLE 6.5: BETO vs. MTL_3 predictions samples from NewsCom-TOX dataset, showing improved MTL performance.

MTL model is the *constructiveness* which shows that it is a good indicator to rule out the presence of toxicity.

The second comparison (BETO vs MTL_3emo_imp) is shown in Table 6.6. Specifically, we show 6 examples, namely 3 false negatives and 3 false positives performed by the baseline BETO model. Regarding the false negatives, the instance with ID 10_040 has been classified by the MTL model as *target person* and also has identified a negative emotion which is *sadness*. We assume that this is a challenging post due to its short length and thus BETO needed more context to identify it as *toxic*, but these two features inextricably related to toxicity, in particular *target person*, have helped the MTL to recognize it as *toxic*. The posts with IDs 10_451 and ID 10_373 convey a negative emotion (*anger*), correctly predicted by the MTL model. This emotion is one of the emotions most inextricably related to offensive language. Therefore, we assume it gives a clue to the system to correctly classify the post as *toxic*, although other linguistic phenomena are not identified. Regarding the false positives (instances with ID 14_055, 18_125, and 18_015), they convey a positive emotion (*joy*) correctly predicted by the MTL system. Therefore, we consider that this affective knowledge can be a clue to the MTL to discard the presence of offensive language.

				MTL_3emo_imp	
ID	Text	Gold	BETO	Toxic	Features
10_040	*Aceptamos si a cambio desapareces tú y tu jefe* (We accept if you and your boss disappear in exchange)	Toxic	Toxic	Toxic	target person, sadness
10_451	*PERFECTO PERO A CAMBIO CONTRATO DE TRABAJO OBLIGATORIO YA ESTAMOS HASTA LOS COJ..... DE TRABAJAR PA MANTENER VAGOS* (PERFECT BUT IN EXCHANGE FOR A MANDATORIY WORK CONTRACT, WE ARE FUCKING DONE OF WORKING TO SUPPORT LAZY PEOPLE)	Toxic	Toxic	Toxic	anger
10_373	*Regularizar el que vivan del robo el trapicheo y del esfuerzo y trabajo de los españoles. Iglesias porqué nos odias tanto?* (To regulate the fact that they live off the theft, fraud and the effort and work of the Spanish people. Iglesias, why do you hate us so much?)	Toxic	Toxic	Toxic	anger
14_055	*lo dices como si hubiera que ser Willy Fog para recorrer las 20 millas que hay entre su casa y Kenosha* (You make it sound like you have to be Willy Fog to travel the miles between his home and Kenosha)	Toxic	Toxic	Toxic	joy
18_125	*Es el poder de la paguita mi amigo* (It's the power of the paycheck, my friend)	Toxic	Toxic	Toxic	joy
18_015	*Abrirán las fronteras para meter a 150.000 negros y marroquíes y arreglado* (They will open the borders to bring in blacks and Moroccans and that's it)	Toxic	Toxic	Toxic	joy

TABLE 6.6: BETO vs. MTL_3emo_imp predictions samples from NewsCom-TOX dataset, showing improved MTL_3emo_imp performance.

			MTL3_emo_imp	
ID	Text	Gold	Toxic	Features
18_201	*Prefiero cobrar el paro/subsidio tirao en el salon de mi casa que deslomarme en el campo por el mismo dinero. Uno quiere trabajar, pero no es gilipollas* (I'd rather get my unemployment benefit lying in my living room than go to the countryside... for the same money. I want to work but I am not an asshole)	Toxic	Toxic	anger, impr_lang
18_146	*Vaya gilipollez. Si vieses la de controles que hay en el campo... Te pilla a un africano sin papeles y se te cae el pelo* (What bullshit. If you saw how many controls there are in the countryside. You catch an African without papers and your hair falls out)	Toxic	Toxic	anger,impr_lang
18_118	*Justo lo contrario a racista. Ahora que la mayoría de los inmigrantes se han vuelto a sus países, se ha destapado que los españoles no quieres hacer ese curro y faltan 150000 personas. Los españoles prefieren quedarse en casa cobrando paguitas* (Just the opposite of racist. Now that most of the immigrants have gone back to their countries it has been revealed that Spanish do not want to do that job and there is a lack of people. The Spanish prefer to stay at home and get a little pocket money)	Toxic	Toxic	anger, target group
10_354	*NO NO NO NO NO NO ESPAÑOLES DESPERTAR TENEMOS QUE DESENTERRAR EL HACHA DE GUERRA NO NO NO NO NO NO* (NO NO NO NO NO NO NO NO NO NO SPANISH AWAKENING WE HAVE TO UNEARTH THE HATCHET NO NO NO NO NO NO NO NO NO NO NO)	Toxic	Toxic	sadness
10_455	*Se dedican a legalizar el insulto y meter morralla en España....FUERA PODEMOS* (They dedicate themselves to legalize insults and to bring trash into Spain. GET OUT PODEMOS!)	Toxic	Toxic	joy
14_236	*Hombre para una vez que la cosa acaba con dos criminales muertos y otro herido cuando iban a matar a un chaval... hay que aprovechar las veces que los sucesos acabn con un final feliz, que no son muchas* (For once the thing ends with two criminals dead and another wounded when they were going to kill a kid, we must take advantage of the times that the events end with a happy ending, which are not many)	Toxic	Toxic	sadness

TABLE 6.7: Samples mislabeled by the MTL_3emo_imp model. Three false positives and three false negatives, respectively.

These examples indicate that our MTL models along with the best linguistic phenomena identified predict the toxic language more accurately than BETO (which does not contain any external knowledge) and is particularly improved in cases that have been missed by the plain model (which is also reflected by the increased recall shown in the results).

6.1.5 Error analysis

To further investigate the proposed MTL model that achieves the best performance, we conducted a qualitative error analysis. Specifically, we mainly analyzed the instances in the test set that were wrongly labeled by the MTL_3emo_imp model in the NewsCom-TOX dataset in order to identify the possible challenges that the system faces with

Spanish toxicity detection. In addition, we also looked at the predictions related to emotion classification and the studied concepts. In particular, three false positive and three false negative comments are shown in Table 6.7. In the first false positive (ID 18_201), there is the presence of *improper language* ("gilipollas" - asshole) but without offending anyone because he is referring to himself and is annoyed with the comments that other people make about unemployment benefit, thus, the emotion *anger* is detected. We assume that the model incorrectly predicts this instance as toxic because these two characteristics tend to appear together in toxic language. Similarly, in the second post with ID 18_146, the MTL mislabeled it because of the presence of *improper language* ("gilipollez" - bullshit) and the emotion *anger* detected. However, the post does not contain a direct or indirect offense. In the third false positive (ID 18_118), the user includes the word "racista" (xenophobic), but without offending anyone and also the comment mentions the group ("inmigrantes" - immigrants). Perhaps, the model mislabeled it because the emotion expressed in the comment and detected by the model is *anger* and the immigrant group appears, two factors which, together, are often associated with the presence of toxicity. In the case of the false negative instances, the first post with ID 10_354 is inciting violence subtly and rhetorically, which complicates the toxicity detection task, and therefore the model misclassified the post as *non-toxic*. In the second sample, the model does not understand the paradox and the predicted emotion is *joy* but the user is expressing anger. In the last sample, the model may predict it wrongly because the user is using sarcasm, a phenomenon that is not identified by the model in this case, further, the emotion is misclassified because the user is not expressing *sadness*.

6.1.6 Discussion

We hypothesized that the detection of toxic language could involve other closed linguistic phenomena. Therefore, we analyzed whether a model training simultaneously in the tasks related to the phenomena of *stereotypes*, *target*, *constructiveness*, *mockery*, *sarcasm*, *improper language*, *insults*, *intolerance*, *aggressiveness*, *positive sentiment*, *negative sentiment* and *emotions* is useful for the purpose of toxic language detection. We used corpora labeled for each of the tasks, we explored how to combine these aspects in our model, and also we studied which combination of these concepts could be the most successful.

Our experiments show the benefits of our enrichment method. In particular, we found that the model that achieves the best performance considers the concepts of *target*, *constructiveness*, figurative language (*mockery* and *sarcasm*), *insult*, *improper language* and *emotions* together. In an analysis of results and a detailed transfer knowledge

analysis, we realized that the model is good at recovering posts from the challenge class (*toxic*) in comparison to the baseline BETO.

Finally, in the error analysis we observed that, in particular, the model finds difficulties in detecting cases of mockery and sarcasm, two phenomena that are even difficult for humans to detect.

6.2 Hate speech and offensive content identification

6.2.1 Problem definition

The widespread adoption of social media platforms has made it possible for users to express their opinions easily in a manner that is visible to a huge audience. These platforms provide a large step forward for freedom of expression. At the same time, social media posts can also contain harmful content like hate speech and offensive language (HOF), often eased by the quasi-anonymity on social media platforms. NLP tools play an important role in supporting the moderation process on online platforms. For the evaluation of these identification tools, continuous experimentation with corpora is necessary. The HASOC track (Hate Speech and Offensive Content Identification) is dedicated to encouraging researchers to develop benchmark NLP models for this purpose.

A third shared task series that took place in 2021 for the third time is HASOC (Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages) [14, 26]. In the first edition of the HASOC, in 2019 [26], Hindi, German and English datasets were created for the definition of HOF based on Twitter and Facebook posts. HASOC 2020 introduced two tasks, one on coarse-grained HOF vs. non-HOF language and one which distinguishes hate, offensive language, and profane language for all these languages. HASOC 2021 was extended by a subtask on code-mixed language [33].

We are going to describe our participation [163] in the coarse-grained identification of HOF in English (Subtask 1A), in the 2021 edition of HASOC[2]. This subtask "Subtask 1A: Identifying Hate, offensive and profane content from the post" is a coarse-grained binary classification challenge in which systems are required to classify tweets into two classes:

- (NOT) Non Hate-Offensive: The post does not contain any hate speech, profane, offensive content.
- (HOF) Hate and Offensive: The post contains hate, offensive, and profane content.

[2] `https://hasocfire.github.io/hasoc/2021/`

6.2.2 Methodology

Maybe the most pertinent question arising from our intuition above – namely that HOF detection is related to the tasks of emotion, sentiment, and target classification – is how this intuition can be operationalized as a computational architecture. Generally speaking, this is a *transfer learning* problem, that is, a problem that involves the generalization of models across tasks and/or domains. There are several strategies to address transfer learning problems; see Ruder [149] for a taxonomy. Structurally, our setup falls into the inductive transfer learning category, where we consider different tasks and have labeled data for each. Procedurally, we propose to learn the different tasks simultaneously, which amounts to MTL. For this reason, we follow the approach developed in this doctoral thesis which is explained in Chapter 5.

We build on a standard contextualized embedding setup where the input is represented by the transformer-based encoder BERT [39]. We add four sequence classification heads to the encoder, one for each task, and fine-tune the model on the four tasks in question (binary/multiclass classification tasks). For the sentiment classification task a tweet is categorized into positive and negative categories; emotion classification classifies a tweet into different emotion categories (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *enthusiasm*, *fun*, *hate*, *neutral*, *love*, *boredom*, *relief*, *none*). Different subsets of these categories are considered in this task depending on the emotion corpus that is used to represent the concept of emotion. Target classification categorizes the target of the offense to an *individual*, *group*, to *others* and to be *not mentioned*; and HOF detection classifies a tweet into *HOF* or *non-HOF*. While training, the objective function weights each task equally. At prediction time, for each tweet in the HASOC dataset, four predictions are assigned, one for each task.

6.2.3 Experimental procedure

Our main research question is whether HOF detection can be improved by joint training with sentiment, emotion, and target.

For model selection, we decided to use the dataset provided by the 2019 edition of the HASOC shared task, under the the assumption that the datasets are fundamentally similar (we also experimented with the HASOC 2020 dataset, but the results indicated that this dataset is sampled from a different distribution than the 2021 dataset). During the evaluation phase, we then used the best model configurations we identified on HASOC 2019 to train a model on the HASOC 2021 training data and produce predictions for the HASOC 2021 test set.

Category	**Dataset**	**Annotation**	**Size**	**Source**
Emotion	CrowdFlower*	Ekman's emo.	39,740	CrowdFlower (2016)
	TEC	Ekman's emo.	21,051	Mohammad (2012)
	GroundedEmo.	sadness, joy	2,585	Liu et al. (2017)
	EmoEvent	Ekman's emo, other	7,303	Plaza-del-Arco et al. (2020)
	DailyDialogues	Ekman's emo.	13,118	Li et al. (2017)
	ISEAR	Ekman's emo, shame, guilt	7,665	Scherer (1994)
Sentiment	SemEval 2016*	neg./pos./neutr.	63,192	Mohammad, Saif M. (2017)
HOF	HASOC 2021*	Non, HOF	5,124	HASOC (2021)
Target	OLID*	None, ind., group, other	14,200	OffensEval (2019)

TABLE 6.8: Selection of resources for EA, SA, and offensive target. The data sets that we use in our final experiments are marked with a star*.

The two main remaining model selection decisions are (a), which corpora to use to train the components?; (b), which components to include? In the following, we first provide details on the corpora we considered, addressing (a). We also describe the details of data preprocessing, training regimen, and hyperparameter handling. The results are reported in Section 6.2.4 to address point (b).

We carry out MTL experiments to predict HOF jointly with the concepts of emotion, sentiment, and HOF target. The datasets are listed in Table 6.8. To represent *sentiment* in our MTL experiments, we use the SemEval 2016 Task 6 dataset [164] composed of 4,870 tweets in total. We include the task of *target classification* with the OLID dataset [5], which consists of 14,100 English Tweets. The concept of *HOF* is modeled based on the HASOC 2021 dataset, which provides three sub-tasks. HASOC 2021 Subtask1A contains 5,214 English tweets split into 3,074 tweets in the training set, 769 in the development set, and 1,281 in the test set.[3]

For *emotion detection*, we consider a set of six corpora in the model selection experiment. These are the Crowdflower data[4], the TEC corpus [165], the Grounded Emotions corpus [166], EmoEvent [135], DailyDialogues[167], and ISEAR [168]. Among the available emotion corpora, we chose those because they cover a range of general topics and/or the genre of tweets.

Data Preprocessing. Tweets present numerous challenges in their tokenization, such as user mentions, hashtags, emojis, and misspellings, among others. To address these

[3] `https://hasocfire.github.io/hasoc/2021/call_for_participation.html`
[4] `https://www.crowdflower.com/data/sentiment-analysis-emotion-text/`

challenges, we make use of the ekphrasis Python library[5] [169]. Particularly, we normalize all mentions of URLs, emails, users' mentions, percentages, monetary amounts, time and date expressions, and phone numbers. For example, "@user" is replaced by the token "<user>". We further normalize hashtags and split them into their constituent words. As an example, "#CovidVaccine" is replaced by "Covid Vaccine". Further, we replace emojis with their aliases. For instance, the emoji 😂 is replaced by the token "*:face_with_tears_joy:*" using the emoji Python library[6]. Finally, we replace multiple consecutive spaces with single spaces and replace line breaks with a space.

Training Regimen and hyper-parameters. In the MTL stage, during each epoch, a mini-batch b_t is selected among all 4 tasks, and the model is updated according to the task-specific objective for the task t. This approximately optimizes the sum of all multi-task objectives. As we are dealing with sequence classification tasks, a standard cross-entropy loss function is used as the objective. For hyper-parameter optimization, we split the HASOC 2021 into train (80 %) and validation data (20 %). Afterward, in the evaluation phase, we use the complete training set of HASOC 2021 in order to take advantage of having more labeled data to train our models. For the baseline BERT, we fine-tuned the model for four epochs, the learning rate was set to $4 \cdot 10^{-4}$ and the batch size to 32. For HASOC_sentiment and HASOC_emotion, we fine-tuned the model for three epochs, the learning rate was set to $3 \cdot 10^{-5}$ and $4 \cdot 10^{-5}$ respectively, and the batch size to 32. For HASOC_target, the epochs were set to four, the learning rate to $4 \cdot 10^{-5}$ and the batch size to 16. For HASOC_all, we fine-tuned the model for two epochs, the learning rate was set to $3 \cdot 10^{-4}$ and the batch size to 16. All the configurations used AdamW as optimizer.

6.2.4 Results

In this section, we present the results obtained by the proposal presented in HASOC 2021 English subtask 1. We use the official competition metric macro-average precision, recall and F_1-score as evaluation measures and further report HOF-specific results, as we believe that, for real-world applications, the detection of the concept HOF is more important than non-HOF. The experiments are performed in two phases: the model selection phase and the evaluation phase, which are explained in the following two sections.

Model Selection (HASOC 2019). As described above, we perform the model selection by training our systems on the training set of HASOC 2019 and evaluating them on the corresponding test set. As we hypothesize that the MTL system trained on related

[5] `https://github.com/cbaziotis/ekphrasis`
[6] `https://pypi.org/project/emoji/`

	Macro-Average		
Emotion dataset	P	R	F_1
TEC	0.7583	0.7900	0.7707
Grounded-Emotions	0.7744	0.7738	0.7741
EmoEvent	0.7739	0.7807	0.7772
DailyDialogs	0.7715	0.7865	0.7783
ISEAR	0.7686	**0.7917**	0.7785
CrowdFlower	**0.7981**	0.7778	**0.7870**

TABLE 6.9: MTL results for HOF detection on HASOC 2019 test, varying the emotion dataset. P: precision, R: recall.

		Macro-Average			Class HOF		
Model		P	R	F_1	P	R	F_1
Baseline	BERT	0.775	0.779	0.777	0.66	0.674	0.667
MTL	HASOC_sentiment	0.773	0.789	0.780	0.646	0.708	0.676
	HASOC_emotion	**0.798**	0.778	0.787	**0.712**	0.642	0.675
	HASOC_target	0.778	0.802	0.788	0.648	**0.736**	0.689
	HASOC_all	0.791	**0.807**	**0.799**	0.674	0.733	**0.702**

TABLE 6.10: MTL results for HOF detection on HASOC 2019 test set. P: Precision, R: Recall.

tasks to HOF detection increased the generalization of the model, we decided to use as a baseline the pre-trained language model BERT fine-tuned on the HASOC 2019 corpora to compare the results.

In order to decide which emotion corpora to use for the task of emotion classification in the MTL setting, we test a number of emotion datasets, obtaining the results shown in Table 6.9. These results are on the main task of hate and offensive language detection, but vary the emotion dataset used for MTL. As can be seen, the best performance is obtained by the CrowdFlower dataset, with a substantial margin in terms of Macro-Precision score. This is despite our impression that this dataset is comparably noisy [170]. We believe that what makes the dataset suitable for HOF detection is that it contains a large number of tweets labeled with a wide range of emotion tags, including *hate*. Therefore, we decided to use this emotion dataset in the MTL setting for the final submission of HASOC 2021.

Table 6.10 shows the results of the MTL models including the different auxiliary tasks on the HASOC 2019 test data. The setting HASOC_all refers to the MTL model trained on the combination of all tasks (HOF detection, emotion classification, polarity classification, and offensive target classification). As can be seen, the MTL models surpass the baseline BERT by at least 2 percentage points Macro-F_1. In particular, the MTL

				MTL		
ID	Tweet	Gold	BERT	HOF	Sent.	Emot.	Targ.
107	But Arlene and the extreme unionists do not want that, and they are the Jenga brick stopping the Tory roof collapsing	$\overline{\text{HOF}}$	HOF	$\overline{\text{HOF}}$	neg.	noemo	None
952	I'ts his choice, you can't force him to get served by Muslims	$\overline{\text{HOF}}$	HOF	$\overline{\text{HOF}}$	neg.	noemo	None
506	Sad watching the UK making a total arse of itself	$\overline{\text{HOF}}$	HOF	$\overline{\text{HOF}}$	neg.	sadness	None
4517	When you got average marks in exam... And ur Dad is like... dad.. She is so Beautiful. :- !	$\overline{\text{HOF}}$	HOF	$\overline{\text{HOF}}$	pos.	noemo	None
254	I don't think I have ever disliked anyone more than I dislike you.	HOF	$\overline{\text{HOF}}$	HOF	neg.	sadness	ind.
684	Yet you project the shortcomings of the muslim ruling class on to others, DEFLECTING, DIVERTING AND LYING TO THE MASSES!!!	HOF	$\overline{\text{HOF}}$	HOF	neg.	anger	None
821	Really, sounds like youre inviting open hostilities again. Are you sure your up to this job? Don't want to be rude but you're just not very bright and have a persistent habit of telling lies too.	HOF	$\overline{\text{HOF}}$	HOF	neg.	fear	ind.

TABLE 6.11: BERT vs. MTL predictions samples from HASOC 2019 test set, showing improved MTL performance. neg.: negative sentiment, pos.: positive sentiment, noemo: no emotion, ind.: individual target, None: not target detected

model that obtains the best performance is HASOC_all, followed by HASOC_target, HASOC_emotion and HASOC_sentiment. The performance of HASOC_all increases by 2 points Macro-F_1 over the baseline, with Macro-Precision increasing roughly 1.5 points and Macro-Recall roughly 2.5 points.

Table 6.10 further shows the results of the MTL models on the HOF class in the HASOC 2019 test set. In all MTL systems except HASOC_emotion, the recall improved over the BERT baseline. The highest improvement in terms of this measure is observed in the HASOC_target model, with an increase of 6.2 points. The precision increases by 5.2 points in the HASOC_emotion model. The best run (HASOC_all) outperforms the baseline BERT with a substantial margin (0.702 to 0.667).

Model Analysis. As we aimed to improve HOF detection results by integrating the MTL model with emotion, sentiment, and target datasets, we decided to use the pre-trained language model BERT in HASOC 2019 corpora as a basis and compared the results of both BERT and MTL on HASOC_all models. The comparison of the two systems can be seen in Table 6.11. Specifically, we show 7 examples, namely 4 false

		Macro-Average		
Model		P	R	F_1
Baseline	BERT	0.801	0.796	0.798
MTL	HASOC_sentiment	0.815	0.784	0.795
	HASOC_emotion	0.819	0.799	0.807
	HASOC_target	0.819	**0.802**	**0.809**
	HASOC_all	**0.824**	0.799	**0.809**

TABLE 6.12: MTL results for HOF detection on HASOC 2021 dev set. P: Precision, R: Recall.

positives and 3 false negatives performed by the baseline BERT model. Regarding the false positives, the first two tweets (IDs 107 and 952) are predicted as HOF by the BERT model but MTL correctly classified them as non-HOF, presumably because although the predicted sentiment is negative, the model could neither recognize a negative emotion nor a target to classify it as HOF. Tweet with ID 506 is also correctly predicted by the MTL model as non-HOF, in this case, although the emotion of sadness is negative, we believe that it is not strongly linked to HOF, moreover, the model does not recognize a specific target directed at HOF. The last false positive (tweet ID 4517) expresses a positive sentiment and the model is able to recognize it, thus we suppose that the MTL benefits from this affective knowledge to classify the tweet as non-HOF. Regarding the false negatives, the tweet with ID 254 has been classified by the MTL system as negative sentiment, negative emotion (sadness), and is directed to a person, therefore as these aspects are closely linked to the presence of HOF, we assume that the MTL takes advantage of these aspects to correctly classify the tweet. The next sample, a tweet with ID 684, expresses a negative opinion and an anger emotion, correctly predicted by the MTL, this emotion is one of the emotions most inextricably related to HOF, and together with the negative sentiment could give a clue to the system to correctly classify the tweet as HOF, although the target is not identified. Finally, instance 821 expresses a negative sentiment towards a person, correctly identified by the MTL model. The model predicts fear for this instance – which we would consider a wrong classification. However, even from this classification (fear instead of anger), the MTL model benefits and makes the correct prediction, which was not possible in the plain BERT model.

These examples indicate that our MTL system predicts the class HOF more accurately than BERT and is particularly improved in cases that have been missed by the plain model (which is also reflected by the increased recall on the HASOC 2019 data).

Model Evaluation (HASOC 2021). For evaluation, we use the dataset provided by the organizers of the HASOC 2021 English subtask 1A. First, we want to verify that the MTL models surpass the baseline BERT also in the evaluation setting. We train

		Macro-Average			Class HOF		
Model		P	R	F_1	P	R	F_1
Baseline	BERT	0.802	0.783	0.790	**0.820**	0.886	0.852
MTL	HASOC_sentiment	0.805	0.784	0.792	0.820	0.891	0.854
	HASOC_emotion	0.790	0.762	0.771	0.800	0.892	0.844
	HASOC_target	0.800	0.776	0.785	0.813	0.892	0.851
	HASOC_all	**0.819**	**0.784**	**0.795**	0.812	**0.917**	**0.862**

TABLE 6.13: MTL results for HOF detection on HASOC 2021 test set (IMS-SINAI Team submissions). P: Precision, R: Recall. The official metric is the macro average score.

all models on the HASOC 2021 training set and test them on the dev set of HASOC 2021. The results obtained are shown in Table 6.12. As can be seen, most of the MTL systems except HASOC_sentiment outperform the baseline, which validates our decision to select these models for the final evaluation of HASOC 2021. HASOC_sentiment does improve over the baseline in Macro-Precision but shows a drop in Macro-Recall. One reason might be that the sentiment data that we use is in some relevant characteristics more similar to the data from 2019 than to the data in the 2021 edition of the shared task.

Table 6.13 finally shows the five models that we submitted to the HASOC 2021 Shared Task as team IMS-SINAI, both with the official macro-average evaluation and the class-specific values (which were reported during the submission period by the submission system). We observe that BERT achieves a Macro-F_1 score of 0.790. The MTL models are, in contrast to the HASOC 2019 results, mostly improved in terms of precision, and less consistently in terms of recall. Considering the target classification and emotion classification in MTL models does not show any improvements, however, the sentiment classification does. These results for the separate concepts are contradicting the results on the 2019 data, which is an indicator that either the evaluation or annotation procedures or the data has changed in some relevant property: In the 2019 data, sentiment+HOF is not better than HOF, but emotion+HOF and target+HOF are. In the 2021 data, it is vice versa. However, when combining all concepts of sentiment, emotion, target, and HOF in one model (HASOC_all), we see an improvement that goes above the contribution by the sentiment model alone. Therefore we conclude that the concepts indeed are all helpful for the identification of hate speech and offensive language.

In addition, we report the results for the class HOF in the same table, without averaging them with the class non-HOF. We find this result particularly important, as the practical task of detecting hate speech is more relevant than detecting non-hate speech. The precision values are lower than the recall values, in comparison to the average results. The

recall is particularly increased in the case of the best model configuration (HASOC_all) with 0.917 in comparison to 0.866 to the plain BERT approach. It is noteworthy that all MTL models increase the recall at the cost of precision for the class HOF. This is both important for practical applications to detect hate speech in the world and from a dataset perspective, as most resources have a substantially lower label count of HOF than for other instances.

6.2.5 Error analysis

To gain a better understanding of the MTL model, we conducted an error analysis that looked at the difficulties this model might have in identifying HOF in English texts. We mainly analyzed some instances in the test set of HASOC 2019 that were wrongly labeled by the MTL. Since the gold labels of the HASOC 2021 set were not publicly available at the time of this study, we have not been able to perform this analysis for this subset. Table 6.14 shows 6 instances, namely 3 false positives and 3 false negatives which represent the most common errors performed by the MTL model. In the first false positive with ID 414, there is expletive language in the use of the word "stupid" but there is no offense to a person or group. The system may have misclassified it because the three characteristics that the MTL system identifies are closely related to the presence of offensive language (negative polarity, anger emotion, and target towards an individual). The second false positive with ID 224 mentions different protected groups (Muslim, Christian, Jew, gay) that are often the target of offense, however, it is not possible to affirm with certainty that there is an offense due to the lack of context. In the third tweet with ID 31 it is noted that the language is not English, and therefore, the system is not able to "understand" the tweet. The code mixed language phenomenon has also been frequently observed in the corpus, which makes the task of HOF detection even more difficult to address by the NLP systems. As for false negatives, the tweet with ID 330 is identified by the MTL model with positive polarity and emotion of joy perhaps because of the presence of the words "your party", however, there is an offense in the tweet although it is implied. The following tweet with ID 42 contains mockery, one of the most difficult linguistic phenomena to detect by the system and closely related to the expression of offensive language, however, the MTL predicts this tweet as non-HOF due to the positive polarity and joy emotion identified on the tweet. Finally, the last false negative with ID 630 was not correctly identified as HOF, perhaps because in the training set we observed that the main word ("fanatic") that makes this tweet HOF is not frequently present.

			MTL			
ID	Tweet	Gold	HOF	Sent.	Emot.	Targ.
414	Correction- they are NOT pulling over drivers. It is for pedestrians & cyclists. If they observe you obeying the law, they may approach you to start a conversation & hand out the certificate and it is only being done in Tempe. Still a stupid idea..	$\overline{\text{HOF}}$	HOF	neg.	anger	ind.
224	You can either be Muslim, Christian, Jew OR Gay.	$\overline{\text{HOF}}$	HOF	neg.	noemo	ind.
31	Navika aunty bhi shock mein soch rahi hai yaar yeh toh humse bhi do kadam aagey hai pati patni Starr who	$\overline{\text{HOF}}$	HOF	neg.	joy	None
330	This was expected of you because we know you and your party believes in HALALA !!	HOF	$\overline{\text{HOF}}$	pos.	joy	None
42	Islam is really great For COOKED FLASH they need 'Halal' For FRESH FLASH 'HALALA'	HOF	$\overline{\text{HOF}}$	pos.	joy	None
630	Tell that fanatic not to use any car or buses or aeroplane which use Muslim petrol diesel.	HOF	$\overline{\text{HOF}}$	neg.	noemo	None

TABLE 6.14: Samples mislabeled by the MTL model on the HASOC 2021 test subset. Three false positives and three false negatives, respectively. neg.: negative sentiment, pos.: positive sentiment, noemo: no emotion, ind.: individual target, None: not target detected

6.2.6 Discussion

Most of the research conducted on the detection of hate speech and offensive language has focused on training automatic systems specifically for this task, without considering other phenomena that are arguably correlated with HOF and could therefore be beneficial to recognize this type of phenomenon.

Our study builds on the assumption that the discourse of HOF could involve other affective components (notably emotion and sentiment), and is, by definition, targeted to a person or group. Therefore, in this paper, as part of our participation in the HASOC FIRE 2021 English Subtask1A, we explored if training a model concurrently for all of these tasks (sentiment, emotion and target classification) via MTL is useful for the purpose of HOF detection. We have used corpora labeled for each of the tasks, we have studied how to combine these aspects in our model, and also we have explored which combination of these concepts could be the most successful. Our experiments show the utility of our enrichment method. In particular, we find that the model that

achieves the best performance in the final evaluation considers the concepts of emotion, sentiment, and target together. This improvement is even more clear in the HASOC 2019 data. In an analysis of the results, we have found that the model is good at improving false positives errors performed by BERT. A plausible mechanism here is that positive sentiments and positive emotions are opposite to the general spirit of hate speech and offensive language so that the presence of these indicators permits the model to predict the absence of HOF more accurately.

This is in line with other previous results on MTL amongst multiple related tasks in the field of affective language. As an example, Akhtar et al. [171] has shown that both tasks of sentiment and emotion benefit from each other. Similarly, Chauhan et al. [10] showed an improvement in sarcasm detection when emotion and sentiment are additionally considered. Particularly the latter study is an interesting result that is in line with our work because the sharp and sometimes offending property of sarcasm is shared with hate speech and offensive language. Further, Rajamanickam et al. [172] has already shown that abusive language and emotion prediction benefit from each other in an MTL setup. This also is in line with our result, given that HOF is an umbrella concept that also subsumes abusive language.

Another aspect to study in more detail is based on the observation of substantial differences between the results of the HASOC 2019 and the HASOC 2021 data. Apparently, the improvements of the MTL model are more clear in the 2019 data. This variance in results is an opportunity to study the aspects that influence the performance improvements when considering related concepts.

6.3 Sexism identification in social networks

6.3.1 Problem Definition

Sexism is any discrimination against people based on gender. This discrimination whose predominant target is women is a widespread cultural component, whose basis is the superiority of men over women in different sectors of life, such as work, politics, society, family, and even advertising. This attitude can be found in different areas such as everyday conversations, statements loaded with discriminatory ideology, contempt for the opinions expressed by women, and even embedded in common sayings and expressions. Unfortunately, in today's society, sexism is still very present in contemporary communication, both written and oral, and is increasingly prevalent on the Web [101].

Detecting sexism online is a challenge for social media moderators. For instance, Amnesty International published a report[7] where Twitter is described as a "toxic place" for women. According to this report, "Twitter is a platform where violence and abuse against them flourishes, often with little accountability." This report also mentions that Twitter is failing in its responsibility to respect women's rights online and this could prejudice their freedom of speech. The General Assembly of the United Nations published research on gender justice and free expression[8]. This research claimed the rise of sexism and misogyny and the suppression of women's freedoms. In addition, it calls for social media platforms to create greater awareness and sensitivity to gender issues in their business operations and activities, including through gender training for their program designers, content policy teams, content moderators, fact-checkers, and others.

The serious consequences of this problem, along with the widespread online content, require rapid solutions to control this type of behavior on the web and help human moderators reduce the volume of sexist content. The NLP field plays a crucial role in the development of automatic systems able to detect this content. However, so far, few studies have addressed sexism detection, especially in languages other than English like Spanish. For this reason, it is important to encourage the NLP community to develop solutions to address this task. In this direction, the first evaluation campaign was born in 2021 to promote the development of NLP approaches for detecting sexist content in English and Spanish tweets and gabs [29]. Due to the great success of the task by the NLP community, a second edition was proposed this year [44].

6.3.2 Methodology

In this section, the methodology our team SINAI followed to address Task 1 of EXIST 2022 shared task is described. It consists of a binary classification task in which computational systems have to decide whether a tweet or Gab post contains sexist expressions or behaviors (sexist) or not (non-sexist). This participation is a continuation of our contribution in the first edition of this EXIST 2021 shared task where our team ranked second among the participants with the first approximation of our MTL approach considering emotions and sentiments as phenomena related to offensive language [173].

In this edition, we focused on studying in depth which interactional phenomena of sexism expression, in addition to sentiments and emotions (explored in our participation in the first edition), can aid in the detection of this content. For this aim, we focus on analyzing different linguistic phenomena including sentiments, emotions, sarcasm, irony, the target, insults, and constructiveness. We hypothesize that these related phenomena could help

[7] `https://bit.ly/3MBoBaF`
[8] `https://bit.ly/3wGyU6z`

in the detection of sexism. For instance, the expression of sexism could involve negative sentiments and emotions. At the same time, rhetorical figures such as irony and sarcasm are used to mask a hurtful message or to make fun in terms of gender. Most of the time the sexist comment is directed at a person or group of people, therefore, the target to whom it is directed plays an important role in the message. A common element that is often present in the expression of sexism is the use of insults or swear words. Finally, constructive criticism is a respectful judgment of another person to provide help or a positive view of a specific circumstance. This phenomenon occurs in non-sexist messages and can be an indicator to detect these messages.

In order to include this hypothesis in a computational architecture, we rely on the main model proposed in this doctoral thesis (see Chapter 5: *"Combining linguistic phenomena through a multi-task approach"*) which is an MTL system that is able to take advantage of the shared knowledge across related tasks to predict more accurately the problem.

6.3.3 Experimental procedure

Datasets. The EXIST dataset provided by the organizers in 2022 includes any type of sexist expression or related phenomenon, including descriptive or denunciatory statements when the sexist message is a denunciation or a statement of sexist behavior. This dataset contains both English and Spanish instances which are labeled according to the tasks proposed by the organizers. In this new edition of EXIST challenge, the EXIST 2021 dataset is provided as training data so the complete dataset is composed of 6,977 tweets for training and 3,386 tweets for testing. In addition, we used other corpora corresponding to tasks that could be related to sexism identification including polarity classification (TweetEval [174] for English and InterTASS for Spanish [151]), emotion classification (EmoEvent [135] for and Universal Joy [158] in both languages), offensive language identification in English (OLID) [13], detection of toxicity in comments in Spanish (DETOXIS) [19], Constructive Comments Corpus (C3) [145] in English and the corpus used in Automatic Sarcasm Detection subtask of 2nd FigLang Workshop at ACL 2020 [175].

Dataset preprocessing. The social media datasets which sources are Facebook, Instagram, or Twitter need some data cleaning steps before including the texts in the model since they present a colloquial register. Therefore, the preprocessing steps we apply are as follows:

- URLs and users' mentions are replaced by the tokens URL and USER, respectively.
- Hashtags are unpacked and split into their constituent words.

Phenomenon	EN	ES
Sentiments	TweetEval	InterTASS
Emotions	EmoEvent (EN) & Universal Joy	EmoEvent (ES) & Universal Joy
Sarcasm	Twitter sarcasm	DETOXIS
Target	OLID	DETOXIS
Insults	OLID	DETOXIS
Constructiveness	C3	DETOXIS

TABLE 6.15: Datasets used for each phenomenon in both English (EN) and Spanish (ES) EXIST subsets.

- Elongated words and repeated characters in words are reduced.
- Emojis are converted to their alias.

System settings. As the EXIST dataset is composed of both English and Spanish texts, we split the EXIST dataset into two subsets (EXIST_en) and (EXIST_es). While training the MTL system, we consider each subset separately, thus we develop two different models: one for Spanish and another for English. Regarding the transformer, for EXIST_en subset, we used the BERT model trained on English tweets[9] and for the EXIST_es subset, we opt for BETO [112], a model trained on Spanish texts. The proposed models have been fine-tuned for 2 epochs, with a learning rate of 2e-5 and batch size of 16, the optimization algorithm is Adamw.

6.3.4 Results

During the pre-evaluation phase, we train the model on the training set and then evaluate it on the test set provided by the organizers. For the evaluation phase, we train the model on the training and validation sets, then we evaluate it on the test set.

6.3.4.1 Pre-evaluation phase

In this phase, we analyze different models and choose the best in terms of performance for the final submission. For this aim, we train our systems on the training set of EXIST 2022 and evaluate them on the validation set. As our hypothesis is that the MTL system trained on related linguistic phenomena to sexism identification helps in the detection of this problem, we decided to compare our results by establishing the baseline BERT fine-tuning on the EXIST 2022 corpora. For both English and Spanish subsets, the

[9] `https://huggingface.co/vinai/bertweet-large`
[9] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased`

		Macro-Average			Class sexist		
Model		P	R	F_1	P	R	F_1
Baseline	BETO	0.7886	0.7889	0.7880	0.8158	0.7649	0.7895
MTL	EXIST_emotion	0.8105	0.8109	**0.8101**	0.8341	0.7925	**0.8128**
	EXIST_sentiment	0.8091	0.8091	**0.8079**	0.8410	0.7774	**0.8080**
	EXIST_sarcasm	0.7994	0.7992	**0.7977**	0.8343	0.7622	**0.7966**
	EXIST_insult	0.7954	0.7956	**0.7944**	0.8249	0.7676	**0.7952**
	EXIST_constructiveness	0.7936	0.7932	**0.7917**	0.8296	0.7542	**0.7901**
	EXIST_target_person	0.7854	0.7846	0.7828	0.8238	0.7409	0.7801

TABLE 6.16: MTL results for sexist detection on EXIST 2022 dev set (EXIST_es subset). Results in bold show the models that outperform the baseline in terms of F_1 score. P: Precision, R: Recall.

		Macro-Average			Class sexist		
Model		P	R	F_1	P	R	F_1
Baseline	BERT	0.7964	0.7858	**0.7867**	0.7629	0.8696	0.8128
MTL	EXIST_emotion	0.7928	0.7844	0.7853	0.7658	0.8584	0.8094
	EXIST_sentiment	0.7601	0.7581	0.7586	0.7586	0.7953	0.7766
	EXIST_sarcasm	0.7900	0.7823	0.7832	0.7653	0.8532	0.8069
	EXIST_insult	0.2378	0.5000	0.3223	0	0	0
	EXIST_constructiveness	0.7264	0.7269	0.7259	0.7539	0.7090	0.7308
	EXIST_target	0.7892	0.7851	0.7859	0.7767	0.8351	0.8048

TABLE 6.17: MTL results for sexist detection on EXIST 2022 dev set (EXIST_en subset). P: Precision, R: Recall.

related tasks (phenomena) we have considered along with the corpora used to train the MTL are described in Table 6.15.

The results obtained after fine-tuning the MTL model on each of the related tasks together with the sexism identification task are shown in Tables 6.16 and 6.17 in Spanish and English subsets, respectively. These results are reported on the main task of sexism identification. We use the official competition metric of macro-average precision, recall, and F_1-score as evaluation metrics and further report sexism-specific performance.

For the results on the EXIST_es subset (Table 6.16) we can observe that all the MTL models except EXIST_target_person surpass the baseline BETO in terms of Macro-F_1 and sexist-F_1 scores. In particular, the setting EXIST_emotion achieved the best performance, followed by EXIST_sentiment and EXIST_sarcasm. The performance of EXIST_emotion increases by 2.21 points Macro-F_1 over the baseline, with macro-Precision increasing roughly 2.19 points and macro-recall by 2.02 points. It should be noted that the best model achieved a significant increase of 2.76 points in terms of the sexist-recall score.

Regarding the evaluation of the MTL models in the EXIST_en subset, we observe that they behave differently from the Spanish subset. As can be seen, the baseline BERT is not outperformed by any MTL model. The settings EXIST_emotion and EXIST_target achieve similar performance to the baseline. However, the EXIST_insult and EXIST_constructiveness performance drop considerably.

After comparing the performance of the MTL models in both subsets we can observe that, depending on the language considered, the related tasks (phenomena) that aid in the detection of sexism are different. Therefore, there are two important parameters that should be carefully analyzed when designing an MTL model for this purpose: the selected datasets and the language.

Model Analysis. Our purpose is to observe whether sexism detection improves by integrating the MTL model with other related linguistic phenomena, we decided to perform a qualitative analysis comparing the results of both BERT and the best MTL models (EXIST_emotion and EXIST_sentiment). In this case, we are going to focus on the Spanish language, since, as we have commented above, no improvements in the MTL models have been achieved in the English subset. On the one hand, the comparison of BETO and EXIST_emotion can be seen in Table 6.18. Specifically, we show 6 examples, namely 3 false positives and 3 false negatives performed by the baseline BETO model. Regarding the false positives, the first three tweets (IDs 9898, 10158, 11006) are classified as sexist but EXIST_emotion model correctly predicted them as non-sexist, probably because the predicted emotion is joy, an emotion that is an indicator of non-sexist messages. Regarding the false negatives, tweets with IDs 9874, 10199, and 11237 are correctly classified by the MTL system as anger emotion, thus, we suppose that as this negative emotion is closely linked to the presence of sexism, it helps the model to successfully identify the tweet as sexist, which was not possible in the plain BETO model. On the other hand, the comparison of BETO and EXIST_sentiment is shown in Table 6.19. In particular, 3 false positives and 3 false negatives are depicted. The first three instances with IDs 9774, 9598, and 9954 are correctly predicted by the EXIST_sentiment model, in this case, they express positive sentiments and the model is able to recognize it, thus we suppose that the MTL benefits from this polarity knowledge to classify the tweets as non-sexist. Regarding the false negatives, texts with IDs 9783, 10063, and 10417 express a negative emotion correctly identifies by the MTL system, the expression of sexism involves a negative sentiment, therefore, this knowledge could give a clue to the system to correctly classify the instance as sexist.

These examples suggest that our MTL model predicts the sexist class more accurately than BETO and is particularly improved in cases that have been misclassified by the

			EXIST_emotion	
ID	Text	BETO	Sexist	Emotion
9898	*Super Enamorada De Estas Tremendas Mujeres, @user Son Tan Divinas #VotaCachers #KCA* (Super In Love With These Tremendous Women, @user They're So Divine #VotaCachers #KCA)	sexist	non-sexist	joy
10158	*Hoy le demostré a mi bestie que ser mujer al volante tiene sus beneficios* (Today I proved to my bestie that being a woman behind the wheel has its benefits)	sexist	non-sexist	joy
11006	*Andalucía guapa, gitana, mujer morena, despierta que eres libre, gitana, de tus cadenas, despierta¡ #Andalucia #FelizDiaDeAndalucia* (Andalusia beautiful, gypsy, brown woman, wake up, you are free, gypsy, from your chains, wake up! #Andalusia #HappyAndalusiaDay) #28F	sexist	non-sexist	joy
9874	*@user Ay pero que vieja pelotuda x diorrr!!! RETUITEA UNA CUENTA PARODIAAAAAAA!!! ESTO VOTARON HIJOS DE PUTA, ESTOOOOOO!!!* (@user Ay but what an old asshole x diorrr!!!!! RETWEET AN ACCOUNT PARODYAAAAAAAAAA!!!!! THIS VOTED THIS MOTHERFUCKERS, THISSSSSSS!!!!!)	non-sexist	sexist	anger
10199	*Voy a dejar las cosas claras desde el principio. Creo en la igualdad entre hombres y mujeres, igualdad de derechos y ante la ley. Condeno toda agresión hacia las mujeres pero también las de las mujeres hacia los hombres. Las denuncias falsas existen y la LIVG es la mayor mierda que un estado de derecho pudo crear.* (I will make things clear from the beginning. I believe in equality between men and women, equality of rights and before the law. I condemn all aggression towards women but also women's aggression towards men. False allegations exist and the LIVG is the biggest piece of shit that a state of law could create.)	non-sexist	sexist	anger
11237	*@user No da. Cuando voy a un lugar fino, el cuate que necesita la factura paga y luego devolvemos el importe. (Entre cuates)Las señoritas empoderadas hacen lo mismo, pero si estás con tu PAGAFANTAS quedar en los mismos términos.* (@user does not give. When I go to a fine place, the guy who needs the bill pays and then we return the amount. (Between guys)Empowered ladies do the same, but if you're with your PAGAFANTAS stay on the same terms.)	non-sexist	sexist	anger

TABLE 6.18: Samples mislabeled by the BETO baseline but correctly labeled by the EXIST_emotion model on the EXIST_es subset.

baseline model (which is also reflected by the increased recall of the EXIST_emotion model on the EXIST_es development subset).

			EXIST_sentiment	
ID	Text	BETO	Sexist	Sentiment
9774	*@user @user mujeres como tu, deben estar postulandose a un cargo, no las actrices que solo hacen show, felicidades avi* (@user @user women like you, should be running for office, not the actresses who just do show, congratulations avi)	sexist	non-sexist	positive
9898	*Super Enamorada De Estas Tremendas Mujeres, @user Son Tan Divinas #VotaCachers #KCA* (Super In Love With These Tremendous Women, @user They're So Divine #VotaCachers #KCA)	sexist	non-sexist	positive
9954	*@user Se nota que es candela dios me encanta esa mujer* (@user I can tell she's candela god I love that woman.)	sexist	non-sexist	positive
9783	*¿Censura? Censura la que nos quisieron imponer a las víctimas de acoso y agresión sexual en el ITAM con la carta de confidencialidad. Son tan privilegiades que son insensibles, sin empatía, son escoria.* (Censorship? Censorship that they wanted to impose on the victims of sexual harassment and aggression at ITAM with the letter of confidentiality. They are so privileged that they are insensitive, without empathy, they are scum.) #ITAMSinCensura	non-sexist	sexist	negative
10063	*Dice la Ministra mujer florero de..., no se que de ultraderecha y su discurso de la meritocracia. Normal, donde esté un buen curso* (I don't know what the Minister *florero de ...* says about the ultra-right and her discourse on meritocracy. Normal, where there is a good course)	non-sexist	sexist	negative
10417	*No conoces una mierda del masculinismo, y aún así, quieres decir lo que te sale de los ovarios* (You don't know shit about masculinism, yet you want to say whatever comes out of your ovaries.)	non-sexist	sexist	negative

TABLE 6.19: Samples mislabeled by the BETO baseline but correctly labeled by the EXIST_sentiment model on the EXIST_es subset.

Test set	**Run**	**Acc**	**P**	**R**	**F_1**
ES	1	0.7500	0.7529	0.7509	0,7497
	2	0.7538	0.7559	0.7546	0.7536
	3	0.7500	0.7529	0.7509	0.7497
EN	1	0.8194	0.8148	0.8206	0.8166
	2	0.6312	0.6180	0.6028	0.6013
	3	0.8194	0.8143	0.8181	0.8158

TABLE 6.20: Results in Subtask 1 on the Spanish and English test set of EXIST shared task. Acc: Accuracy, P: Precision, R: Recall.

6.3.4.2 Evaluation phase

In this section, we present the results obtained by the different runs we have explored in Task 1 (sexism identification).

As part of our participation, we present three runs based on the systems reporting the best performance explored during the pre-evaluation phase. Specifically, we chose the three best models for each language (see Table 6.17 and Table 6.17) and combined them from best to worst performance. The models selected for each language and the differences between the three configurations we presented are described in the following:

- **Run 1**. Baseline (EN) + EXIST_emotion (ES).
- **Run 2**. EXIST_target (EN) + EXIST_sentiment (ES).
- **Run 3**. EXIST_emotion (EN) + EXIST_sarcasm (ES).

In Table 6.20 can be seen the official results obtained by our SINAI-TL in the different runs for both Spanish and English. With respect to the Spanish language, the three models present an accuracy score similar, with the second sentiment-based MTL model being slightly higher. For the English language, the first model selected for run 1 (Baseline) continues to achieve the best performance. However, the second run that considers the target is significantly lower in performance compared to the one achieved during the pre-evaluation phase.

Finally, Table 6.21 shows the results obtained by some participants in Task 1. As we can see, our participation ranks fourth among the participants, with the best run resulting from the combination of the baseline model for English and the EXIST_emotion model for Spanish, followed by runs 3 and 2. Therefore, we consider that the MTL model is a successful system that for the Spanish language provides an improvement over the state-of-the-art BETO. Concerning the English language, we observe that is a challenge

to improve the baseline, perhaps because the BERT baseline already shows values above 80% in the different metrics or the datasets selected for the related phenomena are not the most suitable for the sexism identification task.

Ranking	**Team**	**Acc**	**F_1**
1	avacaondata_1	0.7996	0.7978
2	CIMATCOLMEX_1	0.7949	0.7940
3	I2C_1	0.7883	0.7880
4	**SINAI_TL_1**	**0.7845**	**0.7841**
5	**SINAI_TL_3**	**0.7845**	**0.7839**
40	**SINAI_TL_2**	**0.6928**	**0.6882**
44	Majority Class	0.5444	0.3525

TABLE 6.21: Ranking of participants' systems in subtask 1 of EXIST shared task. Acc: Accuracy.

6.3.5 Error analysis

To deepen our understanding of the best-performing MTL model (EXIST_emotion), we conducted an error analysis examining the challenges faced by this model in identifying sexism in Spanish texts. We mainly analyzed some instances in the development set that were wrongly labeled by the EXIST_emotion model on EXIST_es subset. Since the gold labels of the EXIST_es test set are not publicly available, we have not been able to perform this analysis for this subset. Table 6.22 shows 7 instances, namely 4 false positives and 3 false negatives which represent the most common errors performed by the EXIST_emotion model. In the first false positive with ID 9197, there is an offensive generic female expression (*putas mariconas* - faggot whores) that can also be used to refer to men, but as the target is not clear, we assume that the MTL model is based on the gender of the expression to classify the instance as sexist. In this tweet, it should be clarified that in Spanish, the generic masculine is used to refer to both genders and in this atypical case, the female expression is used to give it an even more derogatory tone, therefore, even for humans, it is difficult to get confused about the associated label. The second false positive with ID 9215 includes self-deprecation since the author is calling herself a whore ("guarra"), therefore, the MTL model is not able to recognize that the instance is expressed in first person and misclassified as sexist. The third tweet with ID 10057 contains offensive language which confuses the MTL model and therefore associates it as sexist. This is one of the main challenges faced by NLP-based systems, the difficulty of differentiating between offensive instances that do or do not involve sexism. The tweet with ID 9534 contains a polysemic word whose meaning may be offensive (*perra* - bitch) or may refer to an animal. In this case, it refers to an animal because the breed "mastín" is mentioned, however, as this word is widely used in a

			EXIST_emotion	
ID	Text	Gold	Sexist	Emotion
9197	*Mucho Punch a nazi pero que luego os escondéis como putas mariconas.* (A lot of Punch a Nazi but then you hide like fucking faggots.)	non-sexist	sexist	anger
9215	*@user sos que pocos ahora me siento una guarra* (@user sos that few now I feel like a bitch)	non-sexist	sexist	joy
10057	*Que le den por culo a la rana gustavo esta de polla macho menuda mierda Mastodon is better than* (Fuck this dick dick frog gustavo male what a piece of shit Mastodon is better than)	non-sexist	sexist	anger
9534	*@user Mi perra es colega de un mastín y me flipan. Comp la cuida!* (@user My dog is a mastiff buddy and I love them. Comp takes care of her!)	non-sexist	sexist	anger
9284	*En rele5 a las ocho de la tarde ,una chica con un super escote.* (In rele5 at eight o'clock in the evening, a girl with a super cleavage.)	sexist	non-sexist	joy
10941	*El perfecto ejemplo de mansplaining. Absténgase de comentar mejor. URL* (The perfect example of mansplaining. Refrain from commenting better. URL)	sexist	non-sexist	others
10659	*Pedroche vete a tomar por culo* (Pedroche go fuck yourself)	sexist	non-sexist	anger

TABLE 6.22: Samples mislabeled by the EXIST_emotion model on the EXIST_es subset. Three false positives and three false negatives, respectively.

sexist context, the MTL model misclassified the tweet as sexist. As for the emotions predicted by the MTL model, we can observe that 3 of the 4 false positives are classified as "anger", an emotion that is closely related to the expression of sexism, which may also have confused the system. Regarding the false negatives, the first instance with ID 9284 is apparently a positive tweet (actually the joy emotion is predicted) but it involves a sexist expression (*un super escote* - a super neckline) which is not identified by the MTL model. The next example with ID 10941 contains an URL with a sexist new, which is not possible to identify by the automatic system since it does not have access to examine the content of the URL. Finally, the tweet with ID 10659 contains an offensive expression (*vete a tomar por culo* - go fuck yourself) which is targeted at a famous woman in Spain (Pedroche), however, as this expression does not appear in the training dataset, the MTL model could not identify it as sexist, although the emotion predicted is negative (anger).

Following this analysis, it can be realized that identifying sexism in text is a complex task, even for humans. We can see different challenges faced by the MTL system including identifying self-deprecation, distinguishing between offensive instances that may or may not contain sexism, identifying expressions of sexism missing in the training set, and dealing with specific language peculiarities in Spanish such as the gender of the expressions, among others.

6.3.6 Discussion

We have described our participation as SINAI-TL team in the second edition of the task sEXism Identification in Social neTworks at IberLEF 2022. We have explored whether

different linguistic phenomena that might be related to sexist expression could help to detect this problem. The experiments have been conducted in two languages: English and Spanish. Our results show that there are two important factors to consider while addressing this task in both languages: the linguistic phenomena considered and the datasets selected. For Spanish, we found that taking into account emotions, sentiments, and sarcasm knowledge helps the detection of sexism. For English, the phenomena studied have not shown any improvement over the baseline BERT. We consider that this fact could be related mainly to the datasets chosen. Therefore, it is important to analyze what are the characteristics that should be in line between the datasets considered to study the linguistic phenomena and the sexism dataset, for instance, the source of the text, the number of categories, and the number of comments, among others.

In the error analysis, we realized how complex is the task of automatic sexism identification. In general, the offensive language detection task is challenging, however, in my opinion, as sexism identification is a more specific problem, it involves more factors to be considered while addressing this task. For instance, it is important to recognize the gender to which this action is directed - sexual discrimination is prejudice or discrimination based on sex or gender - and which idioms are common when expressing this phenomenon - they mainly depend on the language addressing. Other challenges that are common and also occur in the detection of offensive language are the identification of self-deprecating messages and idioms that are missing from the training set and thus are often not recognized when predicting new instances.

Chapter 7

Shared task organization

In this chapter, we describe the shared tasks that have been organized in the framework of this doctoral thesis. We present a description of each shared task, the datasets provided, the systems presented by the participants, and the results obtained. Further, for each shared task, we report the main conclusions observed.

7.1 Motivation

As stated in Chapter 2, Section 2.3, different shared tasks have been proposed in recent years to address the offensive language detection problem. However, most of them focused on English, limiting research to other languages such as Spanish. To fill this gap, an important milestone of this doctoral thesis is the organization of different evaluation campaigns for the Spanish language. We took advantage of the different resources generated during this doctoral study (see Chapter 4: *"Resource generation"*) to propose two different shared tasks within the emotion analysis and offensive language research areas of NLP. The former, which has had two editions (Emotion Detection and EmoEvalEs), intends to promote research in Spanish emotion analysis, and the latter focuses on fostering research in Spanish offensive language detection. These shared tasks have been organized within the Iberian Languages Evaluation Forum (IberLEF) collocated with the Society for Natural Language Processing (SEPLN). The number of teams that have participated in the shared tasks demonstrates the NLP community's interest in addressing these problems in Spanish, underlining the importance of organizing these types of challenges to allow researchers to develop their systems and resources and advance on the state of the art in Spanish.

7.2 Emotion Detection

In the research line of emotion analysis, two shared tasks have been organized in the framework of this doctoral thesis.

The first shared task (TASS 2020: Introducing Emotion Detection) within the Task on Semantic Analysis at SEPLN [176] (TASS task within IberLEF 2020 workshop) took place on September 22. In this edition, two subtasks were organized, the continuation of the polarity classification task, reaching its ninth edition, and the innovation of the second task on emotion analysis. This second subtask was proposed as part of this doctoral thesis in order to promote research on emotion analysis in Spanish using the resource described in Chapter 4 (*"EmoEvent: A Multilingual Emotion Corpus based on different Events"*).

This shared task continued its second edition an evaluation campaign named EmoEvalEs (Emotion Detection for Spanish) [177] at IberLEF 2021, within the 37th International Congress of the Spanish Society for Natural Language Processing (SEPLN 2021). The goal of this continuation was to further promote emotion detection and evaluation for Spanish.

In the following sections, we are going to focus on describing these two editions which shared the same objective: emotion analysis in text applying NLP techniques.

7.2.1 Task description

While polarity classification is a well-established task with many standard datasets and well-defined methodologies, emotion detection has received less attention due to its complexity. In fact, it can be considered a further step in the task of polarity classification since it consists of detecting fine-grained emotions in text, not just positive or negative polarity.

The goal of these shared tasks was to classify the main emotion expressed in a tweet as "*neutral or no emotion*" or as one of the six Ekman's basic emotions [137] that best represent the mental state of the user. The emotion categories are listed below, along with some synonyms:

- *Joy*, including serenity and ecstasy.
- *Sadness*, including pensiveness and grief.
- *Anger*, including annoyance and rage.

- *Surprise*, including distraction and amazement.
- *Disgust*, including disinterest, dislike, and loathing.
- *Fear*, including apprehension, anxiety, concern, and terror.
- *Others*, no emotion or neutral.

The first edition of the shared task took place in a general website `http://tass.sepln.org/2020/` while in the second edition we decided to move it to the popular CodaLab platform in order to gain more visibility: `https://competitions.codalab.org/competitions/28682`.

The challenges faced in the emotion detection task are the following:

1. Lack of context: tweets are short (up to 240 characters)
2. Informal language: misspellings, emojis, and onomatopoeias are common
3. Multiclass classification: the dataset is labeled with seven different classes
4. Imbalance dataset: the number of tweets per emotion category does not follow the same distribution

7.2.2 Dataset

The dataset used for both shared tasks is part of one of the main resources for emotion analysis generated in this doctoral thesis (see Chapter 4: *"EmoEvent: A Multilingual Emotion Corpus based on different Events"*). In particular, for these shared tasks we used the Spanish version of this dataset.

With the purpose of providing the dataset to the participants, in both editions, we decided to replace the hashtags with the keyword *HASHTAG* in order to prevent the automatic classifier from relying on hashtags to categorize the emotion associated with a tweet. Moreover, we replaced the user mentions with *@USER* to anonymize mentions to users. Finally, training, development, and test sets were released to the participants. Table 7.1 shows the number of tweets corresponding to each partition by emotion in the first shared task. As can be seen, the emotions more frequent in the subsets are *joy*, *sadness* and *anger* while the unrepresented ones are *surprise*, *disgust* and *fear*. The training, development, and test sets each contained 5,886, 857, and 1,666 tweets, respectively.

Emotion	**Train**	**Dev**	**Test**
Joy	1,270	185	360
Sadness	706	103	200
Anger	600	87	170
Surprise	241	35	68
Disgust	113	16	32
Fear	67	10	19
Others	2,889	421	817
Total	5,886	857	1,666

TABLE 7.1: Distribution of emotions by subset (Train, Development (Dev), Test) in EmoEvent dataset for Task 2.

Emotion	**Training**	**Dev**	**Test**
Joy	1,227	181	354
Sadness	693	104	199
Anger	589	85	168
Surprise	238	35	67
Disgust	111	16	33
Fear	65	9	21
Others	2,800	414	814
Total	5,723	844	1,656

TABLE 7.2: Distribution of emotions by subset (Training, Development (Dev), Test) in EmoEvalEs 2021.

In the second edition, a data curation of the EmoEvent dataset was performed by eliminating some tweets that we observed that were repeated. Moreover, compared to the first edition in TASS 2020 [176], two new features from the EmoEvent dataset were released to the participants: a label to indicate whether the tweet is offensive (OFF) or not (NO) and the event corresponding to the domain associated to the tweet, namely (*WorldBookDay*, *GretaThunberg*, *Venezuela*, *GameOfThrones*, *LaLiga*, or *SpainElection*). Table 7.2 shows the number of tweets by emotion corresponding to each subset in this second edition.

7.2.3 Evaluation measures

For developing their approaches, in both editions, participants received the training and development subsets of the EmoEvent dataset and the test partition was later provided for evaluation. Finally, the submissions of the participants were compared to the gold standard annotations running the script evaluation developed in order to test their methodologies and identify the winner of the challenge.

The metrics used to evaluate the task were the common classification metrics, namely macro-averaged versions of precision, recall, and F_1 scores, being the macro-F_1 the metric used to rank the participants' systems. For the second edition, the accuracy was incorporated being the one used for rating the participants' systems.

7.2.4 Participants and results

Task 2: Emotion Detection

For "Task 2: Emotion Detection", only two teams submitted their systems and results in this edition. The proposals of the two teams are described in detail below:

- **UMUTeam** [178]. This team performed some preprocessing steps on the tweets before incorporating them into the system. These steps include encoding each letter to its lowercase form, fixing misspellings and typos, contracting white-space characters, and removing expressive lengthening (the intentional elongation of letters in a word to emphasize it). The normalized version of each tweet is used as input for extracting specific linguistic features. Therefore, the system adopted by this team is based on the use of linguistic features in combination with word embeddings. As models, the authors selected Convolutional Neural Networks and SVMs with sentence embedding. The system uses UMUTextStats, a self-developed linguistic tool to extract the above-mentioned linguistic features.
- **EliRFUPV** [179] team adapts TWilBERT, a BERT model on Spanish tweets trained on the Twitter domain, based on the motivation that EmoEvent is composed of Spanish tweets. The authors highlighted the advantages of using this model in comparison with the multilingual version of the BERT model since TWilBERT addresses the language and domain dependency because the model is trained specifically in Spanish tweets. The authors compare the adapted system with the Deep Averaging Networks model they established as a baseline.

Table 7.3 shows the final results obtained by the participant systems on the test set. On the one hand, the ELiRF-UPV team obtained the best macro-averaged F1-score of 0.447. They took advantage of BERT by using the TWilBERT model which is trained on Spanish tweets, then they compared the results with a baseline based on Deep Averaging Networks. This team shows that TWilBERT generalizes better than the baseline, obtaining a +1.5 macro-F_1 score in comparison to the baseline, mainly due to an increment of +3.3 in terms of macro-recall score. On the other hand, the UMUTeam achieved a performance of 0.379 in terms of Macro-F_1 score, presenting a system based

Team	P	R	F_1
ELiRF-UPV	0.443	0.450	0.447
UMUTeam	0.420	0.345	0.379

TABLE 7.3: Final raking of Task 2: Emotion detection at IberLEF 2020. P: Precision, R: Recall.

on the combination of linguistic features and word embeddings. This team shows that the use of the combination of both linguistic features and word embeddings performs better than separately. As a classifier, they used the sequential minimal optimization algorithm which is based on SVMs. In conclusion, as can be observed, the participants' scores are low, demonstrating the difficulty of the emotion classification task in Spanish and the necessity to invest efforts and encourage research on this task in the NLP community.

EmoEvalEs: Emotion Detection for Spanish

In the second edition, unlike the first edition of TASS 2020, we observed a remarkable increase in interest in tackling this task. 70 teams registered on the task, 15 submitted results, and 11 presented working notes describing their systems. The following is a summary of the final proposals submitted.

- **GSI-UPM team - 1st [180]**. This group has studied the combination of different features (like TF-IDF, n-grams, sentiment scores, and the provided *event* and *offensiveness* columns) with encodings of a fine-tuned XLM-RoBERTa model. Although the best submission in the competition was their fine-tuned version of the multilingual RoBERTa model (XLM-RoBERTa), the reported scores on the development set are not much higher than those obtained with Logistic Regression over text representations based on provided event and offensive categories along with sentiment analysis scores.

- **BERT4EVER team - 2nd [181]**. The authors adopt the monolingual Spanish BERT to tackle the task (BETO). In addition, they leveraged two augmented strategies to deal with the imbalanced emotion categories in the dataset, namely continual pre-training, and data augmentation. The best result was obtained with the pre-training of BETO on the training set provided by the EmoEvalEs organizers, by ignoring the labels and performing back translation on the three categories with lower proportion: *disgust*, *fear* and *surprise*.

- **Yeti team - 3rd [182]**. This author used back-translation data augmentation technology to solve the problem of data scarcity and data imbalance. Chinese

and English were used as intermediate languages for back-translation. He mainly enhances the fear and disgust categories. The best result was by entering the offensive labels plus tweets text into the BETO-cased model.

- **URJC team - 4th [183]**. The approach to the emotion detection problem proposed by this team was a fine-tuned version of BETO. They tried both, cased and uncased models, and a third tuning reducing by a 30%, the number of samples within the *others* category. The best result was obtained with a voting system over the three trained models.

- **haha team - 5th [184]**. The tweet, the event, and the offensive features are combined as a new text. Then, URLs, white-space characters, and stop words are removed. The author adopts a masked language model technique for data augmentation in order to increase the training set and avoid over-fitting. Experiments were conducted with three cross-language models: BERT, XLM, and XLM-RoBERTa. The best performance was obtained with the XLM-RoBERTa model. The author shows that the technique used for data augmentation increased the generalization of the model.

- **UMU team - 6th [185]**. The authors explore the combination of explainable linguistic features and state-of-the-art transformers. On the one hand, they used the UMUTextStats tool [186, 187] to extract the linguistic features. On the other hand, they used sentence embeddings and word embeddings from fastText, GloVe, and word2vec (although no details on how to compute sentence embeddings were provided) and sentence embeddings from BETO (pre-trained model) and from a fine-tuned BETO version on the EmoEvalEs dataset. Finally, they combine the features using an ensemble model based on the mode. In their analysis, they show the potential of the linguistic features to provide model-agnostic methods for explainability.

- **RETUYT-InCo team - 9th [188]**. They incorporate a diverse set of features to classical machine learning algorithms and traditional neural networks. The final model is LSTM where authors incorporate features from word2vec and BERT embeddings along with a word feature selection by a variance (ANOVA F-value) method. They mentioned that the most difficult emotion categories to classify by the model were *disgust*, *fear* and *surprise*.

- **WSSC team - 10th [189]**. The authors propose a complex architecture that combines BiLSTM encodings with provided offensiveness and event features. Each

of these three sets of features is the input of one or more feed-forward networks, although no clear details are provided. Despite this complexity, the results are not better than attention-based mechanisms reported by other participants.

- **UPC team - 12th [190]**. The authors propose an approach based on a fine-tuned BETO model on pre-processed tweets. They pre-processed the tweets as follows: i) they removed URLs, hashtags, and numbers; ii) they replaced emojis, emoticons, abbreviations, and laughs; iii) they removed punctuation marks, repeated characters, stopwords, and blank spaces; and iv) they lemmatized the text. They concluded that the submitted system is less accurate for detecting the emotion categories with a small number of samples in the dataset: *fear*, *disgust*, and *surprise*.

- **Dong team - 14th [191]**. It presents a combination of different neural networks (XLM-RoBERTa, Transformer encoding layer, TextCNN, and a final linear one). In first place, they pre-processed the tweets by removing punctuation marks, emojis, empty characters, and other special symbols. Then, they passed the input data to XML-RoBERTa model to obtain word vectors with global features of sentences. Then they input the word vector into a Transformer Encoder for secondary feature extraction and then pass the result into a TextCNN network. Finally, they passed the model output to a fully connected layer for classification.

- **Qu team - 15th [192]**. The authors use the XLM-RoBERTa model to extract the features from training samples and then input the acquired word features into the Bi-GRU model to get the emotional features of comments. Finally, they classify the sentiment tendency by the softmax activation function.

Table 7.4 shows the main results of the participants in the EmoEvalEs Shared Task. We received submissions through CodaLab from 15 participants. 11 teams provided their working notes explaining the systems that were developed for the competition. From all submissions, the best scoring system was presented by the GSI-UPM team, followed by BERT4EVER and Yeti. Between the first two participants, it can be observed that the difference in terms of macro-F_1 and Accuracy is minimal. The best team, GSI-UPM, achieved a macro-F_1 score of 0.717028, exploring the combination of different features with a fine-tuned XLM-RoBERTa model. The team ranked in the second position was BERT4EVER, with a macro-F_1 score of 0.711373 which used BETO along with two augmented strategies to enhance the classic fine-tuned model, namely continual pre-training and data augmentation. The third team, Yeti, achieved a macro-F_1 score of

Team	Accuracy	Macro-P	Macro-R	Macro-F_1
GSI-UPM	(1) 0.727657	(1) 0.709411	(1) 0.727657	(1) 0.717028
BERT4EVER	(2) 0.722222	(2) 0.704695	(2) 0.722222	(2) 0.711373
Yeti	(3) 0.712560	(3) 0.704496	(3) 0.712560	(3) 0.705432
URJC-TEAM	(4) 0.702899	(4) 0.692397	(4) 0.702899	(4) 0.696675
haha	(5) 0.692029	(6) 0.679620	(5) 0.692029	(8) 0.663740
UMUTeam	(6) 0.685990	(7) 0.672546	(6) 0.685990	(7) 0.668407
ffm	(7) 0.684179	(5) 0.682765	(7) 0.684179	(5) 0.682487
fazlfrs	(8) 0.682367	(8) 0.664868	(8) 0.682367	(6) 0.668757
RETUYT-InCo	(9) 0.678140	(9) 0.658314	(9) 0.678140	(10) 0.657367
WSSC	(10) 0.675725	(10) 0.657681	(10) 0.675725	(9) 0.661427
job80	(11) 0.668478	(12) 0.652840	(11) 0.668478	(11) 0.646085
UPCTeam	(12) 0.652778	(14) 0.600479	(12) 0.652778	(12) 0.622223
Timen	(13) 0.617754	(15) 0.597877	(13) 0.617754	(13) 0.600217
Dong	(14) 0.536836	(11) 0.653707	(14) 0.536836	(14) 0.557007
qu	(15) 0.449879	(13) 0.618833	(15) 0.449879	(15) 0.446947

TABLE 7.4: EmoEvalEs official ranking by Accuracy (ranking position per metric is shown in parenthesis) at IberLEF 2021. Macro-P: Macro-Precision, Macro-R: Macro-Recall

0.705432, using back translation data augmentation technology to solve the problem of data scarcity and data imbalance, and tried to input the offensive labels and the text of the tweet into the BETO-cased model.

Most of the teams used neural network solutions to address the task. In particular, Transformers are the most widely used model by the participants in two ways: (1) as encoders to obtain contextualized sentence embeddings features from the text, and (2) fine-tuning the pre-trained model on the task of emotion classification. As tweets from the dataset were in Spanish, most of the teams adopted two available pre-trained language models on Spanish corpora, the multilingual XLM-RoBERTa model and the monolingual BETO model.

Only three teams considered offensive and event information in their approaches (GSI-UPM, haha, and WSSC). In most cases, this led to a slight improvement of the system, so both categories seem to retain certain semantic information related to emotions. In general, the enrichment of the learning process with additional data (through data augmentation techniques) or with additional features beyond neural network encodings, provides some insight into the relevance of hybrid methods for determining subjective information from texts. Although end-to-end solutions like BERT-based models are beneficial, additional characteristics are worth exploring, promoting ensemble-based designs.

Focusing on the classification by emotion categories, some participants mentioned the challenge of classifying those emotions whose presence in the dataset is underrepresented. In particular, these emotions are *fear*, *disgust* and *surprise*. Data augmentation by back translation was applied by two of the teams to address the class imbalance. Also, some teams indicated that the systems faced the challenge of distinguishing complementary emotions, for example, the pairs *disgust* and *anger*, *fear* and *sadness* were often confused, a fact that is reflected by their close locations in the two-dimensional models of emotions.

7.2.5 Conclusion

The 2020 edition of TASS introduced as a novelty a new task: "Task 2: Emotion Detection". This task is inextricably linked to the subject of this doctoral thesis and one of the developed resources, EmoEvent, is used to address the emotion classification task in Spanish.

This first edition attracted a total of 14 participants registered, which demonstrates the interest in the task, although only two teams presented their results, possibly because of the COVID-19 pandemic situation. Participants emphasized the difficulty of handling this task in Spanish, as well as the complexity of training a system with an imbalanced corpus of seven categories. As a result, we can see that there is room for improvement in the systems of the participants since the results obtained were very low.

Due to these challenges and the motivation to continue promoting the research of emotion analysis in Spanish, we decided to continue with the second edition "Emotion detection and Evaluation for Spanish Shared Task" (EmoEvalEs) at IberLEF 2021. In this edition, the participation volume was significantly higher. Specifically, EmoEvalEs attracted 70 participants, 15 of them submitted valid predictions and 11 contributed with a description of their systems. As expected, DL approaches constitute the trend in this text classification task. In addition, the combination of linguistic information confirms the benefits of opting for hybrid solutions. Certainly, some of the most interesting challenges that participants faced were class imbalance and how to combine additional features with deep neuronal network encodings, something that was also observed in the first edition, but in this second edition, the teams applied interesting techniques to overcome these challenges.

Although different dataset partitions have been provided this year than last year, there is a clear improvement in performance. The best result reported in macro-F_1 in 2020 was 0.447. Compared to the best macro-F_1 score in this year's edition (0.717). Teams have gained skills in applying deep neural networks and adapting them to specific tasks. Besides, the participation has raised from 2 to 15 teams. We believe that moving the

competition to CodaLab had the additional effect of more visibility, as can be noticed by the fact that five teams are located in China (four of them from Yunnan University), which represents one-third of the total participants.

In future work, we plan to include the English version of the EmoEvent dataset in the competition in order to promote multilingual emotion analysis research and explore how emotions are expressed for each event based on the cultural differences between English and Spanish speakers.

7.3 MeOffendES: offensive language detection in Spanish variants

The third and last shared task proposed in the scope of this doctoral thesis is MeOffendES [28]. It was proposed in 2021 at IberLEF and co-located with the 37th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021). The main purpose of this shared task was to promote research on the detection of offensive language in Spanish variants. The shared task involved four subtasks, the first two correspond to the identification of offensive language categories in generic Spanish texts from different social media platforms, while subtasks 3 and 4 are related to the identification of offensive language targeting the Mexican variant of Spanish. The first two subtasks, in particular, were proposed in the framework of this doctoral thesis and will be explained in-depth in the following sections.

MeOffendes attracted a large number of participants: a total of 69 signed up to participate in the task, 5 submitted official runs on the test data, and 4 submitted system description papers.

7.3.1 Task description

The primary goal of MeOffendES was to encourage and foster the NLP scientific community in developing offensive language detection systems with a focus on Spanish. For this purpose, four subtasks were proposed. The first two place emphasis on identifying offensiveness in generic Spanish using one of the datasets generated in this doctoral thesis named OffendES (see Chapter 4: *"OffendES: A New Corpus in Spanish for Offensive Language Research"*), whereas the last two focus on detecting this phenomenon in the Mexican variant of Spanish. Since the first two are the ones closely related to this doctoral thesis, we will focus on describing them on detailed. For a description of subtasks 3 and 4, please refer to the overview of the shared task [28].

- **Subtask 1: Non-contextual multiclass classification for generic Spanish**. In this subtask, participants had to classify comments into four different categories:
  - **Offensive, target is a person (OFP).** Offensive text targeting a specific individual.
  - **Offensive, the target is a group of people or collective (OFG).** Offensive text targeting a group of people belonging to the same ethnic group, gender or sexual orientation, political ideology, religious belief, or other common characteristics.
  - **Offensive, the target is different from a person or a group (OFO).** Offensive text where the target does not belong to any of the previous categories, e.g., an organization, an event, a place, an issue.
  - **Non-offensive, but with expletive language (NOE).** A text that contains rude words, blasphemes, or swearwords but without the aim of offending, and usually with a positive connotation.
  - **Non-offensive (NO).** Text that is neither offensive nor contains expletive language.

  This task is called non-contextual because no external information about the comment (social network where it was retrieved or target influencer) was provided. Participants could optionally submit confidence values to predictions (as a probability for each class, so they all sum 1.0) for the four considered categories, to evaluate the agreement of predictions with the confidence of ten human annotators.

- **Subtask 2: Contextual multiclass classification for generic Spanish**. In this subtask, the objective was the same as subtask 1 and with the same categories described, but metadata of the comment (information about targeted users and the related social media) was provided to participants. Participants had access to information associated with each comment: social media source where it was retrieved (Instagram, Twitter, Youtube), influencer name with whom the comment is associated (dalas, wismichu, lauraescane, dulceida, windigyrk, jpelirrojo, wildhater, soyunapringada, miare, javioliveira, nauterplay, and nosoymia) and influencer genre (man, woman).

To reach a huge number of NLP researchers worldwide, MeOffendES was implemented on the CodaLab platform, a popular platform for generating and participating in computational research challenges. The MeOffendES shared task website is open to the public: `https://competitions.codalab.org/competitions/28679`.

Class	Training	Dev	Test
NO	13,212	64	9651
NOE	1,235	22	2340
OFP	2,051	10	1404
OFG	212	4	211
Total	16,710	100	13,606

TABLE 7.5: Distribution of categories by subset (Training, Development (Dev), Test) in MeOffendES 2021.

7.3.2 Dataset

For subtasks 1 and 2 of MeOffendES, we released for the first time the dataset OffendES, one of the main linguistic resources for offensive language detection in Spanish generated in this doctoral study (see Chapter 4: *"OffendES: A New Corpus in Spanish for Offensive Language Research"*). This dataset focused on comments posted to the publications of young influencers from well-known social platforms (Twitter, Instagram, and YouTube) and is manually labeled on offensive pre-defined categories (OFP, OFG, NOE, NO). A subset of the corpus is labeled with three annotators while another subset is labeled with ten annotators. The latter attaches a degree of confidence to each label computed as the ratio of annotators that agreed on the majority label over the total number of annotators, allowing for multiclass classification and multioutput regression research. For the competition, we selected 30,416 posts from the total comments and different sets were released to the participants. During the pre-evaluation phase, training and development (Dev) sets were provided to the participants and for the evaluation phase, the test set was released. Table 7.5 shows the distribution of the comments by category in each subset.

7.3.3 Evaluation measures

For the evaluation of subtasks 1 and 2, we considered the standard classification metrics of micro-averaged and macro-averaged precision, recall, and F_1 measures. In cases where participants submit confidence values (between 0 and 1) to their outputs, we used one of the most preferred metrics for regression tasks, the mean squared error (MSE), a risk metric corresponding to the expected value of the squared (quadratic) error or loss:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{7.1}$$

7.3.4 Participants and results

7.3.4.1 Subtask 1

This subtask, as introduced previously, proposed a pure multi-class text classification problem or a multi-output one. Here, a brief description of the participants' systems is provided.

NLP-CIC team - 1st [93] used the multilingual model XLM-RoBERTa pre-trained on Twitter texts and Sentiment Analysis data. They show that Sentiment Analysis and the social domain adaption are beneficial for the problem of offensive language detection.

UMUTeam - 2nd [193] explored a wide range of features and how to combine them in a final multi-layer perceptron (MLP) with several tentative configurations. The features considered were lexical, negation features, and word and sentence embeddings from different embedding algorithms (fastText, word2vec, gloVe, and a Spanish version of BERT). Word embeddings were evaluated isolated from the rest of the features using convolutional networks and two well-known recurrent architectures like LSTM and Bi-GRU, although MLP was the one showing the best behavior. In general, these features were further pre-processed, with MinMax scaler for linguistic ones and Robust scaler for negation features. All these features were filtered using mutual information. Also, several approaches to combining the total number of features generated were evaluated, including majority voting, weighted voting, and logistic regression. Different kinds of shapes and different numbers of layers, numbers of neurons, dropout probabilities, batch sizes, and activation functions defined a varied number of experiments in order to identify the best configuration for system hyperparameters. From official results, it can be drawn that a combination of BERT-based encodings (pre-trained and fine-tuned), with sentence embeddings and lexical and negation features, became the best solution. When linguistic features were removed, the system obtained the second position in the competition.

The GDUFS_DM team - 3rd applied a sequence classification system fine-tuned on a pre-trained BERT model and composed the final encodings for the text from a max-pooling of the sentence encodings from all layers and token encodings from the last layer. Two additional techniques were integrated into the final system: pseudo-labeling and focal loss. The former technique consists of a two-stage training, where test labels are predicted and re-entered into the learning process to produce a larger training set. The focal loss was used as a way to correct the class imbalance.

Marta Navarrón and Isabel Segura - 4th [194] explored different DL models including LSTM and BERT. The best results were archived by the BERT model. The system ranked in fourth position in the competition.

7.3.4.2 Subtask 2

UMUTeam [193] was the only team submitting results to this subtask. They applied the same system to add to the set of features applied one-hot encodings of contextual columns (gender and media). A robust scaler was also applied to these two features, as done with negation ones. Compared to what was obtained in subtask 1, the integration of contextual information contributed to a small, but consistent improvement in final scores.

7.3.4.3 Results

To evaluate the non-contextual multiclass classification task on the OffendEs dataset, we implemented a straightforward baseline system based on a bag-of-words of unigrams, bigrams, and trigrams and a linear SVM classifier. For the multi-output regression task, we use a multi-output regressor along with the Epsilon-Support Vector Regression. No preprocessing has been applied to the text, nor has a hyperparameter search been performed. We refer to these baselines as *baseline-svm*.

Table 7.6 and Table 7.7 provide a summary of the official results for subtasks 1 and 2 in terms of micro-average and macro-average Precision, Recall, and F_1 scores, respectively. Regarding the multiclass classification setting, it can be noticed that all the teams outperformed our *baseline-svm* which shows the success of the neural network models employed by the participants compared to classical machine learning algorithms. However, for the multi-output regression setting, two of the four teams outperformed the SVM regressor baseline, which shows the success of the classical learning algorithm in this setting. For the non-contextual multiclass classification, it can be seen that the scores of the first three teams are very close. This closeness in performance could be because most of these top-ranked participants relied on similar pre-trained models in their solutions (Spanish BERT model, except for NLP-CIC, who fine-tuned a multilingual RoBERTa model). But greater differences can be observed when looking at the MSE error. The lower the MSE value is, the closer is the system to the behavior of human annotators. In that case, XML-RoBERTa almost reduces to half the error of the second system in the ranking. Finally, both F_1 scores and MSE errors are consistent in terms of ranking order.

For the second subtask, only one team evaluate their system. We can observe that the contextual information did not improve performance, in terms of F_1 score, to that obtained by their system in subtask 1. But regarding MSE, including those additional features (social media platform and gender of the targeted user) do led to a system closer to human annotator behavior.

Subtask 1: Non-contextual classification				
Team	**P**	**R**	**F_1**	**MSE**
NLP-CIC	0.8815	0.8815	0.8815	0.0231
UMUTeam	0.8782	0.8782	0.8782	0.0411
GDUFS_DM	0.8732	0.8732	0.8732	0.0672
Marta_Isabel	0.8416	0.8417	0.8416	0.0697
baseline-svm	0.8285	0.8285	0.8285	0.0615
Subtask 2: Contextual classification				
Team	**P**	**R**	**F_1**	**MSE**
UMUTeam	0.8782	0.8782	0.8782	0.0409

TABLE 7.6: Subtasks 1 and 2 official ranking. Results are in terms of Micro-Precision (P), Micro-Recall (R), and Micro-F_1 scores.

Subtask 1: Non-contextual classification				
Team	**P**	**R**	**F_1**	**MSE**
NLP-CIC	0.7679	0.7093	0.7324	0.0231
UMUTeam	0.7861	0.6919	0.7301	0.0411
GDUFS_DM	0.7565	0.7002	0.7239	0.0672
Marta_Isabel	0.5781	0.5451	0.5595	0.0697
baseline-svm	0.6278	0.4831	0.5236	0.0615
Subtask 2: Contextual classification				
Team	**P**	**R**	**F_1**	**MSE**
UMUTeam	0.7879	0.6921	0.7308	0.0409

TABLE 7.7: Subtasks 1 and 2 official ranking. Results are in terms of Macro-Precision (P), Macro-Recall (R), and Macro-F_1 scores.

7.3.5 Conclusion

MeOffendEs is one of the first shared tasks proposed for the study of offensive language research in Spanish variants. This shared task attempts to continue the research in offensive language detection in Spanish. This shared task attracted a large number of researchers, with a total of 69 participants registered and 4 participating teams in the evaluation phase.

OffendES, a novel dataset on generic Spanish built in this doctoral thesis was used for subtasks 1 and 2, enabling intensive experimentation over a large number of comments from different social media platforms (Twitter, Instagram, Youtube). This evaluation campaign allowed participants to test their systems on this classification task. Different algorithms, features, techniques, and configurations were explored, reporting the effectiveness of state-of-the-art approaches based on pre-trained language models and contributing to the advance of offensive language detection systems. Interesting findings and conclusions have been drawn and very competitive approaches are now available to approach the offensive language detection task in Spanish. Given the great interest from the community, we kept the challenge website open so that anyone interested in trying their methods can do it at any time.

7.4 Closing Remarks

In this chapter, the shared tasks organized in the framework of this doctoral thesis have been presented. They can be split into two main research areas: emotion analysis and offensive language detection. For the former, two editions have been held: the first one in 2020 as a task (Task 2: "Emotion detection") in the TASS 2020 workshop and the second one in 2021 as a shared task: EmoEvalEs. For the latter, two subtasks in the MeOffendES shared task were held in 2021. These shared tasks have taken place in the evaluation forum IberLEF along with the SEPLN conference.

Task 2 at TASS 2020 received only two participants in its first edition, however, a notably increased was observed in the second edition with the organization of EmoEvalEs, as 70 participants register in the task and 11 contributed with a description of their systems. Similarly, the MeOffendES shared task attracted a large number of researchers with a total of 69 registrations and the presence of 4 teams in the evaluation phase of subtasks 1 and 2.

Regarding the approaches presented by the participants, the pre-trained language models constitute the trend in these classification tasks. In addition, the exploration of linguistic features and contextual information and therefore, the development of hybrid solutions have been shown to be promising to address these challenges.

Certainly, some of the most interesting challenges faced by the participants were class imbalance, studying how to combine additional linguistic features in neural networks, lack of context, and the identification of some linguistic phenomena which are involved mainly in subjective messages including irony, sarcasm, or mockery.

It should be noted that the spread of offensive language is a worldwide issue and therefore different languages, cultures, and societies are involved. As a result, we believe that the proposal of these evaluation campaigns, which seek to promote research in languages with limited resources (in this case, Spanish), is essential and will have a positive impact on the advancement and promotion of offensive language detection and emotion analysis in Spanish.

Chapter 8

Conclusion and future directions

With the integration of digital technologies into our daily lives, offensive content has found a way to spread quickly and its regulation is not trivial. As a result, this type of hostile communication can have negative psychological effects on online users, potentially leading to anxiety, bullying, and, in extreme cases, suicide. The seriousness of this problem demands an urgent need for solutions. In this regard, interested stakeholders (governments, online communities, etc.) have proposed possible legislation-based solutions to prevent hostility on the Internet through the implementation of laws and policies. However, these procedures fail to achieve the desired effect because they involve an intensive, time-consuming and costly procedure that limits scalability and quick solutions.

An alternative solution is to rely on NLP-based methods to automatically detect this type of harmful communication. This allows for efficient methods to combat this phenomenon. Due to the great need to develop algorithms of this nature, the offensive language research area has emerged in the NLP field and different efforts have been carried out by the NLP community to foster research in this area. For example, through the organization of different evaluation campaigns, the researchers are asked to conduct different challenges and develop their own solutions based on ML. This type of initiative not only contributes to the development of systems but also to the release of essential datasets for training them.

Deep learning constitutes the state of the art in different NLP tasks, including offensive language detection. In particular, those based on the groundbreaking Transformer architecture are achieving remarkable results. However, although most of the work available so far has successfully applied these systems, at the same time it lacks exploring possible mechanisms for integrating knowledge related to offensive language into the systems. The research covered by this doctoral thesis goes in this direction.

This thesis relies on advanced methods based on transfer learning to tackle the offensive language detection problem. First, we have generated appropriate resources, including corpora and lexicons, to enable us to train ML systems, particularly for Spanish, for which we discovered a significant lack of resources despite it being one of the languages most often used on the Web. Second, we have identified different linguistic phenomena that could occur in the expression of offensive language and proposed a novel methodology that relies on integrating these phenomena into an MTL system to detect more accurate this problem. Furthermore, this methodology has been successfully applied to different scenarios including sexism identification, toxicity detection, and hate speech detection. For each of these scenarios, we have conducted extensive experiments to analyze which linguistic phenomena are most beneficial for the given task. We have also compared the performance of our enriched approach with state-of-the-art methods and found that the integration of this type of knowledge improves the generalization of the model, being able to predict this type of content more accurately.

Finally, the experience in participating in different evaluation campaigns as well as the generation of resources in this doctoral thesis has allowed us to organize different shared tasks. Specifically, in the framework of this doctoral thesis, we have organized two different shared tasks that focus on emotion analysis and offensive language detection. This has provided the NLP community with appropriate resources to develop their own methods and advance in the offensive language research in Spanish.

8.1 Main contributions

The research conducted in this doctoral thesis has resulted in a number of contributions that support the hypothesis outlined in Section 1.2.

To support hypothesis H1, we summarize the following contributions:

(H1) *The subjective nature of offensive language can have strong cultural, demographic, and social implications, and therefore language-specific resources and models are required.*

- We have generated different linguistic resources in the context of offensive language research (Chapter 4). On the one hand, we generated SHARE, a large lexicon of insults and expressions by Spanish speakers. On the other hand, we created three corpora: (1) EmoEvent, a multilingual emotion dataset that allows emotion analysis and offensive language research in both English and Spanish, (2) OffendES, the first large-scale dataset for Spanish offensive language research on three different social media platforms (Youtube, Instagram, and Twitter) allows both classification and regression tasks, and (3) OffendES_span the first Spanish corpus labeled with entities, in which we rely on the SHARE lexicon to expand the OffendES corpus and allow performing NER tasks.
- We have developed our own annotation scheme for each of the resources generated (Chapter 4).
- We have implemented benchmarks for each resource generated in order to evaluate the specific task, validate the resource, and provide the NLP community with preliminary results to compare future approaches (Chapter 4).
- Using the resources generated, we have organized different shared tasks related to emotion analysis and offensive language in order to foster research in these areas for Spanish (Chapter 7).
- We have compared the performance of advanced multilingual and monolingual neural network models based on the Transformer architecture for Spanish offensive language detection (Chapter 3, Section 3.3). We realized that, unlike the multilingual models, the model trained specifically on Spanish texts is able to modulate more accurately Spanish-specific vocabulary and expressions.

To support hypothesis H2, we provide the following contributions:

(H2) *Transfer learning models that leverage linguistic phenomena information related to the expression of offensive language, outperform models for offensive language detection that do not integrate this information.*

- In the literature review (Chapter 2) we found that offensive language research has been addressed mainly as a single optimization task without considering other linguistic phenomena that could be involved in the expression of this behavior.
- We have discussed different phenomena that could be involved in the expression of offensiveness (Chapter 5).
- We have proposed the main methodology conducted in this doctoral thesis which follows an MTL paradigm and relies on integrating the selected linguistic phenomena in a comprehensive computational system for detecting offensive language more accurately (Chapter 5). It relies on approaching different linguistic phenomena involved in the expression of offensive language as tasks in order to simultaneously learn common features among them and improve the generalization of the model. It also benefits from the state-of-the-art pre-trained language models based on the Transformer architecture.
- Using the proposed approach, we obtained performance improvements above the previous state-of-the-art methods (Chapter 5).

The contributions that support hypothesis H3 are summarized as follows:

(H3) *Integrating specific linguistic phenomena into a transfer learning methodology can be beneficial in detecting different offensive scenarios.*

- We have applied the main methodology proposed in this doctoral thesis to different scenarios that involve offensive language research (Chapter 6).
- For each scenario, we have analyzed which linguistic phenomena benefit the task. In addition, we have compared our methodology with state-of-the-art results and conducted an error analysis. Due to an extensive analysis of the results and the errors produced by the methods, we have provided a valuable discussion with the primary findings for each scenario (Chapter 6).
- Using the proposed model, for each scenario, we obtained performance improvements over previous state-of-the-art techniques (Chapter 6).

Finally, after detailing the individual contributions for each hypothesis, the overall contributions resulting from this doctoral thesis can be summarized as follows:

Main contributions

- The generation of different linguistic resources for offensive language research and emotion analysis, focused mainly on Spanish.
- The identification and definition of different linguistic phenomena that could be involved in the expression of the offense.
- The proposal and implementation of an MTL approach based on transfer learning that integrates these phenomena for the detection of offensive language.
- The application of the proposed approach to different scenarios involved in offensive language research.
- The analysis of which linguistic phenomena benefit the most in each scenario through extensive experiments.
- The superior performance of our proposed approach over the previous state-of-the-art approaches.
- The organization of different shared tasks in the IberLEF evaluation campaign using the resources generated in this doctoral thesis to promote offensive language research in Spanish.

8.2 Publications

The different contributions stated in this doctoral thesis have been published in high-impact journals as well as major NLP conferences. Below we are going to present them (in chronological order).

8.2.1 Journals

1. **Plaza-del-Arco, F. M.**, Martín-Valdivia M. T., Jiménez-Zafra S. M., Molina-González, M. D., & Martínez-Cámara, E. (2016). COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. Procesamiento del Lenguaje Natural, 57, 83-90.

   Impact source: SCImago Journal Rankings (SJR): 0.270. **Impact factor**: Q2. **Number of citations (Google Scholar):** 31

2. Jiménez-Zafra, S. M., **Plaza-del-Arco, F. M.**, García-Cumbreras, M. Á., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2018). Monge: Geographic Monitor of Diseases. Procesamiento del Lenguaje Natural, 61, 193-196.

   Impact source: SCImago Journal Rankings (SJR): 0.270. **Impact factor**: Q2. **Number of citations (Google Scholar):** 1

3. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Jiménez-Zafra, S. M., & Martín-Valdivia, M. T. (2018). Lexicon Adaptation for Spanish Emotion Mining. Procesamiento del Lenguaje Natural, 61, 117-124.

   Impact source: SCImago Journal Rankings (SJR): 0.270. **Impact factor**: Q2. **Number of citations (Google Scholar):** 10

4. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). Detecting misogyny and xenophobia in Spanish tweets using language technologies. ACM Transactions on Internet Technology (TOIT), 20(2), 1-19.

   Impact source: WOS (JCR). **Impact factor:** Q2.
   Number of citations (Google Scholar): 54

5. **Plaza-del-Arco, F. M.**, Martín-Valdivia, M. T., Ureña-López, L. A., & Mitkov, R. (2020). Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. Future Generation Computer Systems, 110, 1000-1008.

   Impact source: WOS (JCR). **Impact factor:** Q1.
   Number of citations (Google Scholar): 33

6. López-Úbeda, P., **Plaza-del-Arco, F. M.**, Díaz-Galiano, M. C., & Martín-Valdivia, M. T. (2021). NECOS: An annotated corpus to identify constructive news comments in Spanish. Procesamiento del Lenguaje Natural, 66, 41-51.

   Impact source: SCImago Journal Rankings (SJR): 0.270. **Impact factor**: Q2. **Number of citations (Google Scholar):** 1

7. López-Úbeda, P., **Plaza-del-Arco, F. M.**, Díaz-Galiano, M. C., & Martín-Valdivia, M. T. (2021). How Successful Is Transfer Learning for Detecting Anorexia on Social Media?. Applied Sciences, 11(4), 1838.

   Impact source: WOS (JCR). **Impact factor:** Q2.
   Number of citations (Google Scholar): 4

8. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access, 9, 112478-112489.

Impact source: WOS (JCR). **Impact factor:** Q2.
Number of citations (Google Scholar): 9

9. **Plaza-del-Arco, F. M.**, Casavantes, M., Escalante, H., Martin-Valdivia, M. T., Montejo-Ráez, A., Montes-y-Gómez, M., & Villasenor-Pineda, L. (2021). Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. Procesamiento del Lenguaje Natural.

   Impact source: SCImago Journal Rankings (SJR): 0.270. **Impact factor**: Q2.
   Number of citations (Google Scholar): 20

10. **Plaza-del-Arco, F. M.**, Jiménez Zafra, S. M., Montejo Ráez, A., Molina González, M. D., Ureña López, L. A., & Martín Valdivia, M. T. (2021). Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. Procesamiento del Lenguaje Natural.

    Impact source: SCImago Journal Rankings (SJR): 0.270. **Impact factor**: Q2.
    Number of citations (Google Scholar): 21

11. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120.

    Impact source: WOS (JCR). **Impact factor:** Q1.
    Number of citations (Google Scholar): 51

12. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2022). Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. Knowledge-Based Systems.

    Impact source: WOS (JCR). **Impact factor:** Q1.
    Number of citations (Google Scholar): -

8.2.2 Conferences

1. **Plaza-del-Arco, F. M.**, Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2018). SINAI at IEST 2018: Neural Encoding of Emotional External Knowledge for Emotion Classification. In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 195-200).
   Number of citations (Google Scholar): 2

2. **Plaza-del-Arco, F. M.**, Jiménez-Zafra, S. M., Martín-Valdivia, M. T., & Ureña-López, L. A. (2018). SINAI at SemEval-2018 Task 1: Emotion Recognition in

Tweets. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 128-132).
Number of citations (Google Scholar): 4

3. **Plaza-del-Arco, F. M.**, Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A. (2018). SINAI en TASS 2018: Inserción de Conocimiento Emocional Externo a un Clasificador Lineal de Emociones. Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN) 2018.
Number of citations (Google Scholar): 2

4. **Plaza-del-Arco, F. M.**, Jiménez-Zafra S. M., Martín-Valdivia M. T. , & Ureña-López, L. A. (2018). Using Facebook Reactions to Recognize Emotion in Political Domain. In XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018): Avances en Inteligencia Artificial. pp. 955-960). Asociación Española para la Inteligencia Artificial (AEPIA).
Number of citations (Google Scholar): -

5. Úbeda, P. L., **Plaza-del-Arco, F. M.**, Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2019). Detecting Anorexia in Spanish Tweets. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 655-663).
Number of citations (Google Scholar): 5

6. Puertas, E., Moreno-Sandoval, L. G., **Plaza-del-Arco, F. M.**, Alvarado-Valencia, J. A., Pomares-Quimbaya, A., & Ureña-López, L. A. (2019). Bots and Gender Profiling on Twitter using Sociolinguistic Features. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum.
Number of citations (Google Scholar): 7

7. Moreno-Sandoval, L. G., Puertas, E., **Plaza-del-Arco, F. M.**, Pomares-Quimbaya, A., Alvarado-Valencia, J. A., & Ureña-López, L. A. (2019). Celebrity Profiling on Twitter using Sociolinguistic Features. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum.
Number of citations (Google Scholar): 5

8. **Plaza-del-Arco, F. M.**, López-Úbeda, P., Diaz-Galiano, M. C., Ureña-López, L. A., & Martin-Valdivia, M. T. (2019). Integrating UMLS for Early Detection of Sings of Anorexia. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum.
Number of citations (Google Scholar): 3

9. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Martín-Valdivia, M. T., & Ureña-López, L. A. (2019). SINAI at SemEval-2019 Task 3: Using affective features for emotion classification in textual conversations. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 307-311).
Number of citations (Google Scholar): 3

10. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Martin-Valdivia, M. T., & Ureña-López, L. A. (2019). SINAI at SemEval-2019 Task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 735-738).
Number of citations (Google Scholar): 11

11. Molina-González, M. D., **Plaza-del-Arco, F. M.**, Martín-Valdivia, M. T., & Ureña-López, L. A. (2019). Ensemble Learning to Detect Aggressiveness in Mexican Spanish Tweets. In Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing.
Number of citations (Google Scholar): 8

12. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Martín-Valdivia, M. T., & Ureña-López, L. A. (2019). SINAI at SemEval-2019 Task 5: Ensemble learning to detect hate speech against inmigrants and women in English and Spanish tweets. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 476-479).
Number of citations (Google Scholar): 6

13. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). SINAI at SemEval-2020 Task 12: Offensive language identification exploring transfer learning models. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1622-1627).
Number of citations (Google Scholar): 3

14. **Plaza-del-Arco, F. M.**, Strapparava, C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). EmoEvent: A Multilingual Emotion Corpus based on different Events. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 1492-1498).
Number of citations (Google Scholar): 42

15. García-Vega, M., Díaz-Galiano, M. C., García-Cumbreras, M. Á., **Plaza-del-Arco, F. M.**, Montejo-Ráez, A., Jiménez-Zafra S. M., ... & Moctezuma, D.

(2020). Overview of TASS 2020: Introducing Emotion Detection. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020).
Number of citations (Google Scholar): 18

16. **Plaza-del-Arco, F. M.**, Montejo-Ráez, A., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). OffendES: A New Corpus in Spanish for Offensive Language Research. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 1096-1108).
Number of citations (Google Scholar): 3

17. **Plaza-del-Arco, F. M.**, López-Úbeda, P., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). SINAI at SemEval-2021 Task 5: Combining Embeddings in a BiLSTM-CRF model for Toxic Spans Detection. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 984-989).
Number of citations (Google Scholar): -

18. **Plaza-del-Arco, F. M.**, Halat, S., Padó, S., & Klinger, R. (2021). Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. In Proceedings of the Forum for Information Retrieval Evaluation (FIRE 2021).
Number of citations (Google Scholar): 13

19. Mesa-Murgado, J. A., **Plaza-del-Arco, F. M.**, López-Ubeda, P., & Martín-Valdivia, M. T. (2021). A Social Monitor for Detecting Inappropriate Behavior. In Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021).
Number of citations (Google Scholar): -

20. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Ureña-López L. A.,& Martín-Valdivia, M. T. (2021). SINAI at IberLEF-2021 DETOXIS task: Exploring Features as Tasks in a Multi-task Learning Approach to Detecting Toxic Comments. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021).
Number of citations (Google Scholar): 1

21. **Plaza-del-Arco, F. M.**, Molina-González, M. D., Alfonso, L. & Martín-Valdivia, M. T. (2021). Sexism Identification in Social Networks using a Multi-Task Learning

System. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021).
Number of citations (Google Scholar): 2

22. **Plaza-del-Arco, F. M.**, Molina-González, M. D., & Ureña-López L.A. & Martín-Valdivia M. T. (2022). Exploring the Use of Different Linguistic Phenomena for Sexism Identification in Social Networks. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022).
Number of citations (Google Scholar): -

23. **Plaza-del-Arco, F. M.**, Parras Portillo, A. B., López-Úbeda, P., Botella Gil, B., & Martín-Valdivia, M. T. (2022). SHARE: A Lexicon of Harmful Expressions by Spanish Speakers. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 1307-1316).
Number of citations (Google Scholar): -

24. Mesa-Murgado, J. A., **Plaza-del-Arco, F. M.**, Collado-Montañez, J., Ureña-López, L. A., & Martín-Valdivia, M. T. (2022). ALIADA: Artificial Intelligence-based language applications for the detection of aggressiveness in social networks. In Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022).
Number of citations (Google Scholar): -

25. **Plaza-del-Arco, F. M.**, Collado-Montañez, J., Ureña-López L. A., & Martín-Valdivia, M. T. (2022). Empathy and Distress Prediction using Transformer Multi-output Regression and Emotion Analysis with an Ensemble of Supervised and Zero-Shot Learning Models. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 239-244).
Number of citations (Google Scholar): -

26. Mármol-Romero, A. M., Jiménez-Zafra, S. M., **Plaza-del-Arco, F. M.**, Molina-González, M. D., Martín-Valdivia, M. T., & Montejo-Ráez, A. (2022). SINAI at eRisk@ CLEF 2022: Approaching Early Detection of Gambling and Eating Disorders with Natural Language Processing. In Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum.
Number of citations (Google Scholar): -

27. **Plaza-del-Arco, F. M.**, Martín-Valdivia, M. T., Klinger, R. (2022). Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora. In Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022).
    Number of citations (Google Scholar): 1

8.3 Future directions

From a broader perspective, this doctoral thesis makes various contributions to the offensive language research area using NLP techniques. Although we focused on creating resources and methods for offensive language detection and its scenarios, a number of challenges remain open.

In Chapter 2 we have mentioned the lack of consensus among researchers for defining offensive language and related concepts like HS or abusive language. Although some taxonomies have been proposed by different researchers, there is not a clear taxonomy to follow. We believe that the definitions of these concepts should be clear especially for tasks such as the design of an annotation guide where instructions are provided to annotators on how to annotate these phenomena in the data. Therefore, we believe it is crucial to take this step and encourage interdisciplinary collaboration among domain experts like sociologists, lawyers, and psychologists in order to provide a broad definition.

Another aspect that is perhaps closely linked to the previous point is the difficulty of annotating this type of phenomenon by human annotators. We have observed this aspect especially when creating the resources (Chapter 4). As we have discussed throughout the thesis, offensive language has a major subjective component as it is highly dependent on language, culture, and demographics. Therefore, what may sound offensive to one annotator may not be to another. We believe that in order to address the offensive language detection task, the context of each annotator should be further analyzed and solutions should be designed to integrate this type of knowledge into the NLP systems to mimic the way humans approach this type of task. Recent NLP research is moving in the direction of what is called "learning from disagreement," i.e., developing learning techniques that take into account numerous annotator judgments, as opposed to assuming that there is a single (gold) interpretation for every instance.

We believe our proposed approach (Chapter 5) to integrating different linguistic phenomena through an MTL system offers a new perspective on how to approach the offensive language detection problem. In this thesis, we tested it on a variety of linguistic phenomena we identified closely related to the expression of offensiveness. Further work can

continue to identify other phenomena that can contribute to the detection of this behavior. Similarly, when implementing our system, the objective function has given equal importance to each linguistic phenomenon, but it may be the case that some linguistic phenomena contribute more to the detection of offensive language. Therefore, future work can measure the importance of each phenomenon to the task and incorporate this knowledge into the system, for example, by adapting the objective function with specific weights for each phenomenon.

Finally, recent advances in NLP are showing the success of pre-trained language models as zero-shot learners. Zero-shot learning aims at performing predictions without having seen labeled training examples specific to the concrete task. This opens new directions in how to approach NLP tasks without the need of having label training data. Motivated by the popular GPT-3, prompting has emerged as a viable alternative input format for NLP models to act as zero-shot learners or few-shot learners. Different approaches are currently leveraging prompt methods showing great success. These techniques make it possible to add expert knowledge to model training in a new way that goes beyond manually labeling instances or designing labeling functions. We believe the offensive language research area can benefit from this new strategy and we are convinced that part of future research will go in this direction. It would be interesting to analyze how the instructions are formulated in the prompts to detect offensive language within the text and the ability of these methods to understand these instructions and identify the presence of this behavior.

Appendix A

EmoEvent annotation guidelines

The annotation guidelines created for the generation of two of the corpora (EmoEvent and OffendES) described in Section 3 are shown in detail below.

A.1 The task

- Goal of the task: to label the main emotion expressed in the text.
- Level of anotation: at the document level, i.e., each comment extracted from the Twitter social network is labeled.

A.2 Categories

Each post will be categorized according to one of Ekman's basic emotions or the category *other* shown below:

- **Anger**. This emotion arises when we are blocked from achieving a goal and/or treated unfairly. At its most extreme, anger can be one of the most dangerous emotions because of its potential connection to violence. Synonyms: annoyance, rage.
  - UEFA should be ashamed of this referee tonight. Awful decisions all the time.
  - Don't be a fucking idiot! The world needs your support, udumass!
  - If Arya doesn't end up on that iron throne, I'm gonna be so pissed off!!

- **Disgust**. It contains a series of states with varying intensities ranging from mild disgust to intense repulsion. All disgust states are triggered by the sensation that something is aversive, repulsive and/or toxic. Synonyms: disinterest, dislike, loathing.
  - In case you didn't already know. SOCIALISM SUCKS #Venezuela.
  - When You Think You're Woke But You're Actually Just An Ignorant, Unpleasant Misanthrope.
  - USER Simply put, you're a disgusting man, no discussion and no doubt. Go away;
- **Fear**: Fear arises with the threat of harm, either physical, emotional, or psychological, real or imagined. Synonyms: apprehension, anxiety, terror.
  - I'm nervous. I can't control my emotions while waiting for the final decision.
  - I'm worried about this situation.
  - The movie is terrific! Turn the computer off.
- **Joy**: It is a positive emotion that is usually accompanied by well-being and joy. It is generated as a result of a positive event. Synonyms: serenity, ecstasy.
  - Books are packed with knowledge, insights into a happy life, life lessons, love. I love them.
  - Well done, Greta, for promoting this important message.
  - We are champions of la liga Yesssss. Congratulations!
- **Sadness**. This emotion is a kind of emotional pain or affective state caused by spiritual decay and often expressed by weeping, a downcast face, lack of appetite, lassitude, etc. A person may feel sad when his/her expectations are not met or when life circumstances are more painful than joyful. The opposite emotion is joy. Synonyms: pensiveness, grief.
  - Watching such a beautiful and historically important building burned to the ground broke my heart.
  - #NotreDameCathedralFire is sad, but it's even more sad that so much of the world only acknowledges the tragedies of the white/wealthy.
  - We are with you all in this time. My thoughts are with you :(.
- **Surprise**. This is defined as a reaction caused by something that is unexpected, strange or novel to the person. Synonyms: distraction, amazement.
  - Best free kick I've ever seen. WOW. The guy isn't human.

  - Absolutely shocking. I can't believe it.
  - Wow! Even Satan mourns at this great loss!

- **Other**. When the comment does not imply any relationship with the emotions or arouses an emotion other than those described above.
  - Engineering experts consider fire damages and reconstruction questions.
  - Nike Air Max 90 Essential Midnight Navy UK Size 8 Mens Trainers Free.
  - Investigators believe the cause of the tragedy is an electrical issue.

In addition, each post will be classified as offensive or non-offensive according to the following definition:

The text is offensive if it contains some form of unacceptable language (blasphemy). This category includes insults, threats or bad words.

A.3 FAQ

- What if the text is objective, i.e., it does not express any emotion? In this case, the post will be considered as *Other*.
- What if more than one emotion is present in the text? In this case, the main emotion expressed in the text is labeled.
- What if the text is also offensive? In this case, you will label the text as offensive in addition to the emotion expressed in the post.

A.4 Important notes

This task is about emotion expressed in texts, not about the emotion you may feel when reading it from a personal point of view. *Please try to be as objective as possible.*

Please read the instructions provided for labeling carefully and thoroughly. **Note that it is only possible to choose one of the categories defined for each comment**.

It is important that you do not overthink about the answer, **follow your first intuition**.

A.5 Remember

The commentary should be annotated by **considering only the information contained in the text**, without thinking about what happened before or after the fact or situation expressed in the text.

If you are in doubt about which category should be selected, **the vocabulary used in the commentary may help you decide on the most appropriate category**.

Appendix B

OffendES annotation guidelines

B.1 The task

This annotation guide aims to provide a set of instructions to perform the labeling of a corpus of comments extracted from three social networks: Twitter, Youtube and Instagram. Specifically, it is required to label the offensiveness in these comments that are responses given by different users to publications of certain influencers in these social network platforms.

- Goal of the task: to label offensiveness in social media comments.
- Level of annotation: at the document level, i.e., each comment extracted from the social networks (Twitter, Youtube, and Instagram) is labeled.

B.2 Categories

Before describing each of the categories to be labeled, it is important to read the definition of when a comment is considered offensive:

A comment is considered offensive when language is used to make an explicitly or implicitly directed offense that may include insults, threats, messages containing profane language, or profanity.

Below are the categories to be labeled along with detailed examples. Specifically, three offensive categories and two non-offensive categories:

- **Offensive, the target is a person (OFP)**. Offensive text targeting a specific individual.

  - ¡Vete a tu casa!, eres un borracho.
  - No te tocaría ni con un palo, me das asco.
  - Nunca voy a llegar a entender a este tipo, no suelta más que estupideces por su boca cada vez que habla.

- **Offensive, the target is a group of people or collective (OFG)**. Offensive text targeting a group of people belonging to the same ethnic group, gender or sexual orientation, political ideology, religious belief, or other common characteristics.

  - Las feminazis deberían de aprender antes a fregar que a manifestarse.
  - Estoy harto de lidiar con estos listillos de izquierdas que apoyan ideas en contra a sus ideales.
  - El día de los homosexuales no debería de existir, al final van a tener más derechos que nosotros.

- **Offensive, the target is different from a person or a group (OFO)**. Offensive text where the target does not belong to any of the previous categories, e.g., an organization, an event, a place, an issue.

  - ¡Joder! Este estúpido confinamiento va a acabar con mi paciencia, estoy harto de estar encerrado.
  - Vaya mierda de manifestación hicieron ayer, espero que no se vuelva a perder el tiempo y dinero en hacer este tipo de chorradas.
  - Por mi parte, ya pueden cancelar todos los estúpidos e inútiles festivales de este año, me es totalmente indiferente.

- **Non-offensive, but with expletive language (NOE)**. A text that contains£ rude words, blasphemes, or swearwords but without the aim of offending, and usually with a positive connotation.

  - ¡Eres el puto amo! Me encanta tu vídeo.
  - Joder, ¡qué mierda!, voy a tener que volver a repetir todo el ejercicio de nuevo.
  - ¡Me cago en dios, qué dolor me ha dado en el dedo del pie pequeño al darme contra la mesa!

- **Non-offensive (NO)**. Text that is neither offensive nor contains expletive language.

  - ¡Buenos días amigos! ¿Qué tal va el fin de semana?
  - No me ha gustado nada ver esa serie de miedo, han sido dos horas larguísimas.

– No entiendo por qué siempre se tienen que quejar por lo mismo, a mí no me parece tan mal la idea.

B.3 FAQ

- If more than one directed category appears in a comment, the order of prevalence would be as follows: offensive directed at a person, offensive directed at a group, offensive directed at others. For example: What a shitty political party (offensive to a group) I can't stand your president (offensive to a person). In this case, you would choose offensive directed at a person.
- What if the message contains bad language but conveys something positive? For example, What a bastard you are! You got the job you were hoping for. Congratulations! In this case the non-offensive category with bad language would be chosen.
- What if not a single swear word is used but it is clearly offending an individual/group/others? In this case the comment is offensive and you would choose the target (individual, group or others) to whom it is directed.
- What happens if the offensive comment is directed at a place, city or country? For example, it is not worth living in this mean and ruinous Spain that is not worth a penny. In this case you would choose the offensive category directed at others.

B.4 Important notes

This task is about offensiveness in texts, not about the offensiveness you may feel when reading it from a personal point of view. *Please try to be as objective as possible.*

Please read the instructions provided for labeling carefully and thoroughly. **Note that it is only possible to choose one of the categories defined for each comment**.

It is important that you do not overthink about the answer, **follow your first intuition**.

B.5 Remember

The commentary should be annotated by **considering only the information contained in the text**, without thinking about what happened before or after the fact or situation expressed in the text.

If you are in doubt about which category should be selected, **the vocabulary used in the commentary may help you decide on the most appropriate category**.

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[2] Carmen Aguilera-Carnerero and Abdul Halik Azeez. 'Islamonausea, not Islamophobia': The many faces of cyber hate speech. *Journal of Arab & Muslim media research*, 9(1):21–40, 2016.

[3] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221, 2010.

[4] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*, 2017.

[5] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1144.

[6] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060.

[7] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47, 2020.

[8] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[9] Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179, January 2021. ISSN 1588-2861. doi: 10.1007/s11192-020-03737-6.

[10] Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.401.

[11] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online, June 2021. Association for Computational Linguistics.

[12] Paula Fortuna and Sérgio Nunes. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4), jul 2018. ISSN 0360-0300. doi: 10.1145/3232676.

[13] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010.

[14] Thomas Mandl, Sandip Modha, Gautm Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 12th Forum for Information Retrieval Evaluation*, 2020.

[15] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive Language Detection in Online User Content. In *Proceedings of the 25th*

International Conference on World Wide Web, WWW '16, page 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883062.

[16] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Detecting aggressors and bullies on Twitter. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 767–768, 2017.

[17] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80, 2012. doi: 10.1109/SocialCom-PASSAT.2012.55.

[18] General Policy Recommendation no. 15 of the European Commission. *Hate Speech*. 2016. URL `http://hudoc.ecri.coe.int/eng?i=REC-15-2016-015-ENG`.

[19] Mariona Taulé, Alejandro Ariza, Montserrat Nofre, Enrique Amigó, and Paolo Rosso. Overview of the DETOXIS Task at IberLEF-2021: DEtection of TOXicity in comments In Spanish. *Procesamiento del Lenguaje Natural*, 67, 2021.

[20] Cambridge Dictionary. *Profanity*. 2017. URL `https://dictionary.cambridge.org/dictionary/english/profanity`.

[21] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[22] S.V. Kogilavani, S. Malliga, K.R. Jaiabinaya, M. Malini, and M. Manisha Kokila. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, 2021. ISSN 2214-7853. doi: https://doi.org/10.1016/j.matpr.2021.04.102.

[23] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, Vienna, Austria, 2018.

[24] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language*

Processing (KONVENS 2019), pages 354–365, Erlangen, Germany, 2019. German Society for Computational Linguistics & Language Technology.

[25] Elisabetta Fersini, Paolo Rosso, and Mary E. Anzovino. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN*, 2018.

[26] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450377508. doi: 10.1145/3368567.3368584.

[27] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007.

[28] Flor Miriam Plaza-del-Arco, Marco Casavantes, Hugo Jair Escalante, M. Teresa Martín-Valdivia, Arturo Montejo-Ráez, Manuel Montes-y-Gómez, Horacio Jarquín-Vásquez, and Luis Villaseñor-Pineda. Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67, 2021. ISSN 1989-7553.

[29] Francisco Ródriguez-Sánchez, Jorge Carrillo de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67(0), 2021. ISSN 1989-7553.

[30] Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://aclanthology.org/W18-4400`.

[31] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[32] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[33] Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 1–3, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450395960. doi: 10.1145/3503162.3503176.

[34] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19(21):4654, 2019.

[35] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[36] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, oct 2003. ISSN 1046-8188.

[37] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain, April 2017. Association for Computational Linguistics.

[38] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[40] Sabit Hassan, Younes Samih, Hamdy Mubarak, and Ahmed Abdelali. ALT at SemEval-2020 Task 12: Arabic and English Offensive Language Identification in Social Media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1891–1897, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[41] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.430.

[42] Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.629.

[43] Çağrı Çöltekin. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

[44] Francisco Rodríguez-Sánchez, Jorge Carrillo de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of EXIST 2022: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 69(0), 2022. ISSN 1989-7553.

[45] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1095.

[46] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1044.

[47] Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS, 2018. URL `http://ceur-ws.org/Vol-2253/paper49.pdf`.

[48] Tullio De Mauro. Le parole per ferire. *Internazionale*, 27(9):2016, 2016. URL `https://www.internazionale.it/opinione/tullio-de-mauro/2016/09/27/razzismo-parole-ferire`.

[49] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302, 2002. URL `http://multiwordnet.fbk.eu/paper/MWN-India-published.pdf`.

[50] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250, 2012. URL `https://doi.org/10.1016/j.artint.2012.07.001`.

[51] Anna Schmidt and Michael Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101.

[52] Cagatay Catal, Ugur Sevim, and Banu Diri. Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm. *Expert Systems with Applications*, 38(3):2347–2353, 2011.

[53] Andrew McCallum, Kamal Nigam, et al. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[54] Rodrigo Moraes, JoãO Francisco Valiati, and Wilson P GaviãO Neto. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2):621–633, 2013.

[55] Mikalai Tsytsarau and Themis Palpanas. Survey on Mining Subjective Data on the Web. *Data Mining and Knowledge Discovery*, 24(3):478–514, 2012.

[56] Nello Cristianini, John Shawe-Taylor, et al. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge university press, 2000.

[57] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS, 2018.

[58] Shervin Malmasi and Marcos Zampieri. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202, 2018.

[59] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3: 1137–1155, March 2003.

[60] Ronan Collobert and Jason Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

[61] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[62] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[63] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. ISSN 0364-0213. URL https://www.sciencedirect.com/science/article/pii/036402139090002E.

[64] Alex Graves. Generating Sequences With Recurrent Neural Networks. *CoRR*, abs/1308.0850, 2013. URL http://arxiv.org/abs/1308.0850.

[65] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

655–665, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1062. URL https://www.aclweb.org/anthology/P14-1062.

[66] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://www.aclweb.org/anthology/D14-1181.

[67] Björn Gambäck and Utpal Kumar Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, 2017.

[68] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742, 2018.

[69] Gustavo Henrique Paetzold, Marcos Zampieri, and Shervin Malmasi. UTFPR at SemEval-2019 task 5: Hate speech identification with recurrent neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 519–523, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2093. URL https://www.aclweb.org/anthology/S19-2093.

[70] I Goenaga, A Atutxa, K Gojenola, A Casillas, A Dıaz de Ilarraza, N Ezeiza, M Oronoz, A Pérez, and O Perez de Vinaspre. Automatic Misogyny Identification Using Neural Networks. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain*, 2018.

[71] Alison Ribeiro and Nádia Silva. INF-HatEval at SemEval-2019 Task 5: Convolutional Neural Networks for Hate Speech Detection Against Women and Immigrants on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425, 2019.

[72] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Trans. Internet Technol.*, 20(2), mar 2020. ISSN 1533-5399. doi: 10.1145/3377323. URL https://doi.org/10.1145/3377323.

[73] Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067, 2019.

[74] Ping Liu, Wen Li, and Liang Zou. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2011. URL `https://aclanthology.org/S19-2011`.

[75] Abigail S Gertner, John Henderson, Elizabeth Merkhofer, Amy Marsh, Ben Wellner, and Guido Zarrella. MITRE at SemEval-2019 Task 5: Transfer Learning for Multilingual Hate Speech Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 453–459, 2019.

[76] Alon Rozental and Dadi Biton. Amobee at SemEval-2019 Tasks 5 and 6: Multiple Choice CNN Over Contextual Embedding. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 377–381, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[77] Miriam Benballa, Sebastien Collet, and Romain Picot-Clemente. Saagie at Semeval-2019 Task 5: From Universal Text Embeddings and Classical Features to Domain-specific Text Classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 469–475, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2083. URL `https://www.aclweb.org/anthology/S19-2083`.

[78] Hajung Sohn and Hyunju Lee. MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations. *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559, 2019.

[79] Tharindu Ranasinghe and Marcos Zampieri. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.470. URL `https://aclanthology.org/2020.emnlp-main.470`.

[80] Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. fBERT: A Neural Transformer for Identifying Offensive Content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.findings-emnlp.154`.

[81] Ricardo Martins, José Almeida, Pedro Henriques, and Paulo Novais. Increasing Authorship Identification Through Emotional Analysis. In *World Conference on Information Systems and Technologies*, pages 763–772. Springer, 2018.

[82] Flor Miriam Plaza-del-Arco, M Dolores Molina-González, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez. SINAI at SemEval-2019 Task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 735–738, 2019.

[83] Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, and Pedro Henriques. Hate Speech Classification in Social Media Using Emotional Analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66, 2018. doi: 10.1109/BRACIS.2018.00019.

[84] Axel Rodríguez, Carlos Argueta, and Yi-Ling Chen. Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174. IEEE, 2019.

[85] Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. Attending the Emotions to Detect Online Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online, November 2020. Association for Computational Linguistics.

[86] AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. Leveraging Affective Bidirectional Transformers for Offensive Language Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 102–108, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL `https://aclanthology.org/2020.osact-1.17`.

[87] Ibrahim Abu Farha and Walid Magdy. Multitask Learning for Arabic Offensive Language and Hate-Speech Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, 2020.

[88] Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Joint Modelling of Emotion and Abusive Language Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.394.

[89] Simona Frenda, Ghanem Bilal, et al. Exploration of Misogyny in Spanish and English tweets. In *Third workshop on evaluation of human language technologies for iberian languages (ibereval 2018)*, volume 2150, pages 260–267. Ceur Workshop Proceedings, 2018.

[90] Mario Graff, Sabino Miranda-Jiménez, Eric Sadit Tellez, Daniela Moctezuma, Vladimir Salgado, José Ortiz-Bejar, and Claudia N Sánchez. INGEOTEC at MEX-A3T: Author Profiling and Aggressiveness Analysis in Twitter Using μTC and EvoMSA. In *IberEval@ SEPLN*, pages 128–133, 2018.

[91] Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org, 2018.

[92] Diego Benito, Oscar Araque, and Carlos A. Iglesias. GSI-UPM at SemEval-2019 Task 5: Semantic Similarity and Word Embeddings for Multilingual Detection of Hate Speech Against Immigrants and Women on Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 396–403, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2070.

[93] Segun Taofeek Aroyehun and Alexander Gelbukh. Evaluation of Intermediate Pretraining for the Detection of Offensive Language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org, 2021.

[94] Flor-Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña López, and M. Teresa Martín-Valdivia. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Technol.*, 20(2), March 2020. ISSN 1533-5399. doi: 10.1145/3369869. URL `https://doi.org/10.1145/3369869`.

[95] Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021. ISSN 0957-4174.

[96] Zeerak Waseem and Dirk Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. URL `https://aclanthology.org/N16-2013`.

[97] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118, 2018.

[98] Rachel Noelle Simons. Addressing Gender-Based Harassment in Social Media: A Call to Action. *iConference 2015 Proceedings*, 2015.

[99] Linda Beckman, Curt Hagquist, and Lisa Hellström. Discrepant gender patterns for cyberbullying and traditional bullying – An analysis of Swedish adolescent data. *Computers in Human Behavior*, 29(5):1896–1903, 2013.

[100] Jesse Fox and Wai Yen Tang. Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. *Computers in Human Behavior*, 33:314–320, 2014.

[101] Jesse Fox, Carlos Cruz, and Ji Young Lee. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52:436–442, 2015. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2015.06.024.

[102] Cristina Bosco, Viviana Patti, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Francesco Ruffo, Rossano Schifanella, and Marco Stranisci. Tools and Resources for Detecting Hate and Prejudice against Immigrants in Social Media. In *Symposium III. Social Interactions in Complex Intelligent Systems (SICIS) at AISB 2017*, pages 79–84. AISB, 2017.

[103] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine Learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[104] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[105] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference for Learning Representations, San Diego, 2015*, 2015. URL `http://arxiv.org/abs/1412.6980`.

[106] Cristian Cardellino. Spanish Billion Word Corpus and Embeddings. *Retrieved from*, 2016. URL `https://crscardellino.github.io/SBWCE/`.

[107] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[108] M Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, and José M Perea-Ortega. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257, 2013.

[109] Juan Manuel Pérez and Franco M. Luque. Atalaya at SemEval 2019 Task 5: Robust Embeddings for Tweet Classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 64–69, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2008.

[110] Luis Enrique Argota Vega, Jorge Carlos Reyes-Magaña, Helena Gómez-Adorno, and Gemma Bel-Enguix. MineriaUNAM at SemEval-2019 task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 447–452, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[111] Abigail Gertner, John Henderson, Elizabeth Merkhofer, Amy Marsh, Ben Wellner, and Guido Zarrella. MITRE at SemEval-2019 Task 5: Transfer Learning for Multilingual Hate Speech Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 453–459, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[112] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*, 2020.

[113] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P16-1162`.

[114] Yoshua Bengio. Practical Recommendations for Gradient-Based Training of Deep Architectures. In *Neural Networks: Tricks of the Trade*, 2012.

[115] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

[116] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[117] Luis Enrique Argota Vega, Jorge Carlos Reyes-Magaña, Helena Gómez-Adorno, and Gemma Bel-Enguix. MineriaUNAM at SemEval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 447–452, 2019.

[118] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. doi: https://doi.org/10.1177/001316446002000104.

[119] Ron Artstein and Massimo Poesio. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL `https://aclanthology.org/J08-4004`.

[120] Beatriz Botella-Gil, Flor Miriam Plaza-del-Arco, Ana Belén Parras Portillo, and Yoan Gutiérrez. Fiero: Asistente virtual para la captación de insultos. *Procesamiento del Lenguaje Natural*, 2021. URL `http://ceur-ws.org/Vol-2968/paper8.pdf`.

[121] Pancracio Celdrán. *El gran libro de los insultos: tesoro crítico, etimológico e histórico de los insultos españoles*. La Esfera de los Libros, 2009.

[122] Nuria Gala and Mathieu Lafourcade. NLP lexicons: innovative constructions and usages for machines and humans. In *eLEX'2011: Electronic LEXicography in the 21st century: new applications for new users*, page 12, 2010. URL `https://hal-lirmm.ccsd.cnrs.fr/lirmm-00832983`.

[123] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165 (1):91–134, 2005. doi: https://doi.org/10.1016/j.artint.2005.03.001.

[124] Antonio Toral and Rafael Muñoz. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, 2006. URL `https://aclanthology.org/W06-2809`.

[125] Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.752. URL `https://aclanthology.org/2020.acl-main.752`.

[126] Flor Miriam Plaza-del-Arco, Arturo Montejo-Ráez, L. Alfonso Ureña-López, and María-Teresa Martín-Valdivia. OffendES: A New Corpus in Spanish for Offensive Language Research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, Held Online, September 2021. INCOMA Ltd.

[127] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977. URL `https://doi.org/10.2307/2529310`.

[128] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001, 2001.

[129] Saif M Mohammad. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Tex. In *Emotion measurement*, pages 201–237. Elsevier, 2016.

[130] The YouTube Team. More updates on our actions related to the safety of minors on YouTube. `http://web.archive.org/web/20080207010024`, 2019. Accessed: 2020-01-10.

[131] Jenn Chen. 2020 Social media demographics for marketers. `https://sproutsocial.com/insights/new-social-media-demographics/`, 2020. Accessed: 2020-09-22.

[132] Philip M McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

[133] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[134] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online, October 2020. Association for Computational Linguistics.

[135] Flor Miriam Plaza-del-Arco, Carlo Strapparava, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

[136] Flor Miriam Plaza-del-Arco, Ana Belén Parras Portillo, Pilar López-Úbeda, Beatriz Botella-Gil, and María Teresa Martín-Valdivia. SHARE: A Lexicon of Harmful Expressions by Spanish Speakers. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1307–1316, Marseille, France, June 2022. European Language Resources Association. ISBN 979-10-95546-34-4.

[137] Paul Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, 6(3-4): 169–200, 1992. doi: 10.1080/02699939208411068.

[138] Robert Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001. URL `https://www.jstor.org/stable/27857503`.

[139] G. T. W. Patrick. The Psychology of Profanity. *Psychological Review*, 8(2):113–127, 1901. doi: 10.1037/h0074772.

[140] Wafa Alorainy, Pete Burnap, Han Liu, Amir Javed, and Matthew L. Williams. Suspended Accounts: A Source of Tweets with Disgust and Anger Emotions for Augmenting Hate Speech Data Sample. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 581–586, 2018. doi: 10.1109/ICMLC.2018.8527001.

[141] Axel Rodríguez, Carlos Argueta, and Yi-Ling Chen. Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174, 2019. doi: 10.1109/ICAIIC.2019.8669073.

[142] Bo Pang, Lillian Lee, et al. Foundations and Trends® in Information Retrieval. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, 2008.

[143] Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[144] Varada Kolhatkar and Maite Taboada. Constructive Language in News Comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3002. URL `https://aclanthology.org/W17-3002`.

[145] Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. Classifying Constructive Comments. *CoRR*, abs/2004.05476, 2020.

[146] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly Abusive Language–What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, 2021.

[147] Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

[148] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[149] Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, NUI Galway, 2019. URL `https://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf`.

[150] Rich Caruana. Multitask Learning. *Machine learning*, 28(1):41–75, 1997.

[151] Eugenio Martínez-Cámara, Manuel C Díaz-Galiano, Miguel A García-Cumbreras, Manuel García-Vega, and Julio Villena-Román. Overview of TASS 2017. *Proceedings of TASS*, pages 13–21, 2017.

[152] Eugenio Martínez Cámara, Yudivián Almeida-Cruz, Manuel Carlos Díaz-Galiano, Suilan Estévez-Velarde, Miguel Ángel García Cumbreras, Manuel García Vega, Yoan Gutiérrez, Arturo Montejo-Ráez, Andrés Montoyo, Rafael Muñoz, Alejandro

Piad-Morffis, and Julio Villena-Román. Overview of TASS 2018: Opinions, Health and Emotions. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018*, volume 2172 of *CEUR Workshop Proceedings*, pages 13–27. CEUR-WS.org, 2018.

[153] Manuel Carlos Díaz-Galiano, Manuel García Vega, Edgar Casasola, Luis Chiruzzo, Miguel Ángel García Cumbreras, Eugenio Martínez Cámara, Daniela Moctezuma, Arturo Montejo-Ráez, Marco Antonio Sobrevilla Cabezudo, Eric Sadit Tellez, et al. Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. In *IberLEF@ SEPLN*, pages 550–560, 2019.

[154] Mario Graff, Sabino Miranda-Jiménez, Eric Sadit Tellez, Daniela Moctezuma, Vladimir Salgado, José Ortiz-Bejar, and Claudia N. Sánchez. INGEOTEC at MEX-A3T: Author Profiling and Aggressiveness Analysis in Twitter Using μTC and EvoMSA. In *IberEval@SEPLN*, 2018.

[155] M Guzman-Silverio, A Balderas-Paredes, and AP López-Monroy. Transformers and Data Augmentation for Aggressiveness Detection in Mexican Spanish. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*, pages 293–302, 2020.

[156] Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. Detecting Aggressiveness in Mexican Spanish Social Media Content by Fine-Tuning Transformer-Based Models. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*, pages 236–245, 2020.

[157] Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. SINAI at IberLEF-2021 DETOXIS task: Exploring Features as Tasks in a Multi-task Learning Approach to Detecting Toxic Comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 580–590. CEUR-WS.org, 2021.

[158] Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. Universal Joy A Data Set and Results for Classifying Emotions Across Languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, 2021.

[159] Christopher Zimmerman, Mari-Klara Stein, Daniel Hardt, and Ravi Vatrapu. Emergence of Things Felt: Harnessing the Semantic Space of Facebook Feeling

Tags. In *Thirty Sixth International Conference on Information Systems, Fort Worth*, 2015.

[160] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating Mutual Information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138.

[161] Guillem García Subies. GuillemGSubies at IberLEF-2021 DETOXIS task: Detecting Toxicity with Spanish BERT. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 591–598. CEUR-WS.org, 2021.

[162] Angel Felipe Magnossão de Paula and Ipek Baris Schlicht. AI-UPV at IberLEF-2021 DETOXIS task: Toxicity Detection in Immigration-RelatedWeb News Comments Using Transformers and Statistical Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 547–566. CEUR-WS.org, 2021.

[163] Flor Miriam Plaza-del-Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. In *FIRE 2021 Working Notes*, pages 297–318, 2021. URL `http://ceur-ws.org/Vol-3159/T1-30.pdf`.

[164] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics.

[165] Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[166] Vicki Liu, Carmen Banea, and Rada Mihalcea. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483, 2017. doi: 10.1109/ACII.2017.8273642.

[167] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[168] Klaus R. Scherer and Harald G. Wallbott. The ISEAR Questionnaire and Codebook. Geneva Emotion Research Group, 1997.

[169] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[170] Laura-Ana-Maria Bostan and Roman Klinger. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[171] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1034.

[172] Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Joint Modelling of Emotion and Abusive Language Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.394.

[173] Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, Luis Alfonso Ureña López, and María Teresa Martín-Valdivia. Sexism Identification in Social Networks using a Multi-Task Learning System. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 491–499. CEUR-WS.org, 2021.

[174] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088.

[175] Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. A Report on the 2020 Sarcasm Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.1.

[176] Manuel García Vega, Manuel Carlos Díaz-Galiano, Miguel Ángel García Cumbreras, Flor Miriam Plaza-del-Arco, Arturo Montejo Ráez, Salud María Jiménez-Zafra, Eugenio Martínez-Cámara, Cesar Antonio Aguilar, Marco Antonio Sobrevilla Cabezudo, Luis Chiruzzo, and Daniela Moctezuma. Overview of TASS 2020: Introducing Emotion Detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN)*, 2020.

[177] Flor Miriam Plaza-del-Arco, Salud María Jiménez-Zafra, Arturo Montejo-Ráez, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67:155–161, 2021.

[178] José Antonio García-Díaz, Angela Álmela, and Rafael Valencia-García. UMUTeam at TASS 2020: Combining Linguistic Features and Machine-learning Models for Sentiment Classification. In *Proceedings of TASS 2020: Workshop on Semantic Analysis at SEPLN (TASS 2020)*, CEUR Workshop Proceedings, Málaga, Spain, September 2020. CEUR-WS.

[179] José Ángel González, Lluís-Felip Hurtado, Ferran Pla, and José Arias Moncho. ELiRF-UPV at TASS 2020: TWilBERT for Sentiment Analysis and Emotion Detection in Spanish tweets. In *Proceedings of TASS 2020: Workshop on Semantic Analysis at SEPLN (TASS 2020)*, CEUR Workshop Proceedings, Málaga, Spain, September 2020. CEUR-WS.

[180] Daniel Vera, Oscar Araque, and Carlos A. Iglesias. GSI-UPM at IberLEF2021: Emotion Analysis of Spanish Tweets by Fine-tuning the XLM-RoBERTa Language Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[181] Yingwen Fu, Ziyu Yang, Nankai Lin, Lianxi Wang, and Feng Chen. Sentiment Analysis for Spanish Tweets based on Continual Pre-training and Data Augmentation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[182] Hongxin Luo. Emotion Detection for Spanish with Data Augmentation and Transformer-Based Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[183] Jorge Alberto Flores Sánchez, Soto Montalvo Herranz, and Raquel Martínez Unanue. URJC-Team at EmoEvalEs 2021: BERT for Emotion Classification in Spanish Tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[184] Kun Li. HAHA at EmoEvalEs 2021: Sentiment Analysis in Spanish Tweets with Cross-lingual Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[185] José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. UMUTeam at EmoEvalEs 2021: Emotion Analysis for Spanish based on Explainable Linguistic Features and Transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[186] José Antonio García-Díaz, Mar Cánovas-García, and Rafael Valencia-García. Ontology-driven Aspect-based Sentiment Analysis classification: An Infodemiological case study regarding infectious diseases in Latin America. *Future Generation Computer Systems*, 112:614–657, 2020. doi: 10.1016/j.future.2020.06.019.

[187] José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506 – 518, 2021. ISSN 0167-739X. doi: 10.1016/j.future.2020.08.032. URL `http://www.sciencedirect.com/science/article/pii/S0167739X20301928`.

[188] Luis Chiruzzo and Aiala Rosá. RETUYT-InCo at EmoEvalEs 2021: Multiclass Emotion Classification in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[189] Fedor Vitiugin and Giorgio Barnabò. Emotion Detection for Spanish by Combining LASER Embeddings, Topic Information, and Offense Features. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[190] Ariadna de Arriba, Marc Oriol, and Xavier Franch. Applying Sentiment Analysis on Spanish Tweets Using BETO. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[191] Suidong Qu, Yanhua Yang, and Quinyu Que. Emotion Classification for Spanish with XLM-RoBERTa and TextCNN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[192] Yuanchi Qu, Shuangjun Jia, and Yanjie Zhang. Sentiment Analysis in Spanish Tweets: The Model based on XLM-RoBERTa and Bi-GRU. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain*, 2021.

[193] José Antonio García-Díaz, Salud María Jiménez-Zafra, and Rafael Valencia-García. UMUTeam at MeOffendEs 2021: Ensemble Learning for Offensive Language Identification using Linguistic Features, Fine-grained Negation and Transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org, 2021.

[194] Marta Navarrón García and Isabel Segura Bedmar. Detecting Offensiveness in Social Network Comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org, 2021.