Are really different topic classification and sentiment analysis?

¿Son realmente diferentes la clasificación temática y el análisis de sentimientos?

Francisco Javier Rufo Mendo
Universidad Nacional de Educación a Distancia
Sierra de Tormantos 21, 10600, Plasencia
frufo3@alumno.uned.es

Resumen: Este trabajo pretende dar una nueva visión a las tareas de clasificación temática de textos y el análisis de sentimientos. A pesar de que se trata de tareas de clasificación, suelen ser abordadas de forma diferente. En el presente trabajo, se lleva a cabo la elaboración de un clasificador que lleve a cabo las dos tareas de forma indiferente, obteniendo resultados similares y comprobando que tal vez pueda existir una única solución para las tareas de clasificación. También presenta un análisis del comportamiento de un clasificador supervisado frente a un clasificador semi-supervisado.
Palabras clave: clasificación temática, tweets, bolsa de palabras, análisis de sentimiento

Abstract: This paper aims to give a new vision to the work of topic classification of text and sentiment analysis. Although it is classification tasks are usually addressed differently. In the present work, we carried out the development of a classifier that performs the two tasks indifferently, with similar results and checking that perhaps there can be a single solution for classification tasks. It also presents an analysis of the behavior of a supervised classifier compared to semi-supervised classifier.
Keywords: topic clasification, tweet, bag of words, sentiment analysis

1 Introduction

Since the early beginning of IR (Information Retrieval) systems, one of its objectives was to determine the topic of the texts analyzed. These texts often contained a large amount of words, so they had a large vocabulary that allowed researchers to develop systems able to perform this task with more or less success.

With the rise of Internet popularity, a lot of web pages, blogs and social networks have been proliferated. The two first cases are considered as normal text documents, so the same techniques can be applied to categorize them. On the other hand, we usually have very short texts when posting in social networks and classic techniques do no fit well enough, thus we are forced to develop new techniques.

We can find an example of social network usage in Twitter[1], where users can express their ideas, opinions, sentiments, etc. in messages called tweets. The main characteristic of tweets is that they have a maximun length of 140 characters. However, users are allowed to insert different elements like URLs, hashtags (any word starting whith "#") or mentions (any username starting with "@") whenever the message length not exceed the limit length.

In 2012 Twitter reached 500 million users, more than 340 millions of daily new messages and 1.6 billions of daily search. Thus, Twitter has become one of the mainly news sources more than a simply social network, just as expressed in Kwak et al. (2010).

The huge amount of information available on Twitter has provoked a lot of researches based on that information. We can find works related with messages categorization

[1] http://www.twitter.com

sentiment analysis, political tendency, event detections, human reaction against natural disasters, users relations and behaviours, etc.

Most of the studies based on Twitter data are related with tweet categorization and sentiment analysis. These two tasks have been always taken as different task, although they both are some kind of classification task, even when their objective can be considered different. These two tasks can be described as follows:

- **Topic classification:** Given a set of predefined classes, each message must be classified into one of them. In some ocassions, a message can be classified into more than one class, these are multiclass systems.
- **Sentiment analysis:** This task tries to determine the sentiment of a message, being able to determine if the message is positive, negative or neutral, or even determine how strong is the sentiment.

In this work we carry out a short survey to tell us how different are these tasks, in order to glimpse the possibility or not to carry out the development of a classification system that serves to address the two tasks at same time, what means that the system should be able to classify tweets and do sentiment analysis without modifing it. At the same time, we try to compare results obtained by supervised and semi-supervised methods.

The structure of the present work is the following. In Section 2 the motivation that has led to address the present work and previous works are described. Section 3 describes the system developed. In section 4 results obtained with the corpus provided by TASS Workshop[2], followed by the conclusions and future work in section 5.

2 Motivation and previous work

This work emerges as a solution for the final work for the subject *Web mining* in *Language and Computer Systems Master Degree*, given by the UNED in Spain. So, this works could not be as professional as others proposed by experts in this area, but pretends to give another point of view by a novel researcher.

This work was mainly focused on topic classification for Spanish language, specifically in the work developed under TASS Workshop. The first edition of TASS Workshop defined tasks for both sentiment analysis and topic classification, for this purpose a corpus with tweets in Spanish language has been provided (Villena-Román, 2012).

A lot of work has been made in both Topic Classification and Sentiment Analysis, but they are mainly focused in English language. Here, we are interested in Spanish langage, where a few works have been proposed. To develop this work, we have based mainly on the works accepted for the first edition of TASS Workshop. Thus we can find different aproaches tested in English and adapted to work in Spanish.

One of the first works for short texts can be found in (Sriram et al., 2010), where the author proposes to take all the messages of the user to determine his class. The main disadvantage is that an user can tweet messages about different topics, and with this aproximation this is not allowed. In this work, all messages will be processed independently.

Although there were not much papers presented at the first edition of TASS, they where very diverse, covering mostly all the tradicional methods for text classification. We can find models based on Maximun Entropy (Ribeiro, 2012) where also special items like URLs, hashtags and mentions are used as characteristics.

While traditional works use only one Bag-of-Words (BOW), in (Martínez-Cámara, 2012) a variety of them are built. They build a BOW using Google Adwords KeyWordTool[3], which given a term returns some related terms. Also they build another BOW with hastags and take into account the number of emoticons in each message for sentiment analysis.

A Latent Dirichlet Allocation (LDA) based method could not be missed. In this way, (Tamara Martín-Wanton, 2012) propose the use of Twitter-LDA (Zhao et al., 2011), although they do not obtain very good results at classifying with the corpus provided by TASS.

A comparation study of a variety of methods can be found in (Santos, 2012), using WEKA (Hall et al., 2009) application. In this work, the use of multiple dictionaries is proposed, given that tweets are similar to SMS messages, and they can have their

[2]http://www.daedalus.es/TASS2013

[3]https://adwords.google.com/select/home

own vocabulary. At the same time they use a spell-checker to correct writing mistakes. The comparation is based on diferent configurations, like using special items allowed in tweets, n-grams, etc.

3 System description

3.1 Data processing

As (Laboreiro et al., 2010) exposed, the first task might be a tokenization of the contents, using spaces between words and punctuation symbols to perform this task. As in (Ribeiro, 2012), repeated chars are eliminated from terms. In this way, only two repetitions of a character are allowed, i.e., the word "holaaaaa" (*"hellooo"* would be replaced by "holaa" (*"helloo"*. Finally, *stop-words* were removed, to avoid terms with high frequency but do not discriminate between classes.

Also different optional task are proposed to process messages content. This task are performed after tokenization process, and they are:

- **Stemming:** Performs a stemming process over the detected terms. For this purpose, Snowball[4] is selected.
- **Process URLs:** If the message contains any URL, its accesed and the content is used as aditional information. Only content under *title, h1 and h2* tags is retrieved, using JSOUP[5] for this task.
- **Hashtags:** Some topics are related to specific hashtags. This option uses hashtags as aditional features.
- **Mentions:** On some occasions, mentions can be a useful feature to determine the topic of the message, because the destination user is related to a specific topic.

Independently of the configuration, URLs, mentions and hashtags are eliminated from the message content to reduce noise and concentrate in message content.

As mentioned, a Bag-of-Words model was developed, but with one peculiarity. Only unique terms for a topic were inserted into de bag of the topic. This is done under the assumption that terms that appear only in messages related to a topic are more discriminant. As with the terms, the same process is made with mentions and hashtags, while URLs terms are treated as normal terms.

3.2 Feature selection

Build a system based BOW approximation was selected to carry out the task of Topic Classification and Sentiment Analysis. As can be expected, the vocabulary obtained from the corpus might be very big, and a lot of terms will be repeated among messages of different topics. To take a first contact, only unique terms will be take into account for each topic. For this purpose, a BOW is built for each topic, and only terms that do not appear in messages from other topics are inserted. Thus, the terms of each topic tend to be more determinant.

Term Frequency (TF) has been used as weighting method, because it is easy to calculate, it is fast, and also because it usually obtains good results.

At this point, noisy terms have been removed, and frequency of unique terms per topic have been calculated. As the number of features is still unmanageable, at least for a novel research like me, only top terms have been selected. In this way, the indicated percent of terms with more frequency have been selected for each topic. Even now, as can be viewed in Table 1 the number of features is high.

3.3 Classifier

In (Santos, 2012), different approaches and classification techniques has been analyzed, in particular all classifiers available in WEKA. Some of the best results were obtained using Complement Naïve Bayes (CNB), having also used Multinomial Naïve Bayes and Sequential Minimal Optimization (SMO) (Platt, 1999).

Bayesians classifiers are statistical and popular learners. Under the assumption that a document can be classified only under one class, the creation of a document is modeled as follows:

1. Each class c has an associated prior probability $Pr(c)$, with $\sum_c Pr(c) = 1$. The author of a document selects a random topic at first.
2. Exists a distribution conditioned by the subject or class $Pr(d|c)$ for each class.

Thus, the total probability of generating a document of class c is $Pr(c)Pr(d)$. Finally,

[4]http://www.snowball.tartarus.org
[5]http://jsoup.org

given the document d, the posterior probability that d was generated from the class c is, using Baye rules:

$$Pr(c|d) = \frac{Pr(c)Pr(d|c)}{\sum_{\gamma} Pr(\gamma)Pr(d|\gamma)} \quad (1)$$

Naïve Bayes classifiers belong to Bayesian family and are widely used due to its simplicity and speed of training. Finally, within the Naïve Bayes classifiers, Naïve Bayes Complement (CNB) is proposed in (Rennie et al., 2003) to deal with biased information, such as the tweets. CNB was proposed as a solution to one of the systemic errors of Naïve Bayes. This error consists in skewed data bias, what means that the existence of more training examples for one class than another can cause the decision boundary to be biased. This fact can cause that the classifier prefers one class over the others in an unwittingly way. While in Multinominal Naïve Bayes uses training data from a class c when estimating weights, CNB uses data from all classes except c. As we will see later, TASS corpus has an irregular distribution of classes examples in training dataset, so a classifier could easily derive into skewed data bias.

Instead of developing a new classifier, WEKA has been used, where among others, CNB is implemented.

4 Evaluation

Even though the work presented for testing in TASS has been made using CNB classifier, other evaluation processes have been made previously using Support Vector Machines (SVM)(Liu, 2006), which obtained worse results.

As some authors noted, maybe TASS corpus is not a very good corpus for training, at least supervised classifiers. This is because it has a bad distributions among topics as can be viewed in Table 2. Other authors relate bad results with a small training dataset comparing to the size of the test dataset, but this case is more real than artificial corpus that have a bigger training dataset than the test dataset.

As we can see in literature (Liu, 2006), in this situation, a co-training model would be more suitable than a supervised one. This is why, in adition to CNB, a co-training version of CNB to TASS 2013 has been presented too.

The main evaluation work has been made with training test. Several experiments have been made and were evaluated with 10-fold cross-validation. The number of terms selected for each experiments were: 20% for content terms, 30% of hashtags and 30% of user mentions.

Configurations and results of this experiments are shown in Table 1. In these experiments, we can see that instead of what we can think, stemming process has lower precision that experiments without it. This may be due to the nature of stemming process. We are working with short texts, and their vocabulary might be very specific. If we stem words, we can find words with different meanings reduced to the same term.

Although almost the same results are obtained, the best configuration is the described for experiment 2, even when the number of features is lower. Surprisingly, the results remains unchanged form all the experiments with SVM classifier, but the precision is much lower than those obtained with CNB.

After examining the results, experiment 2 with CNB was presented to TASS 2013. As said before, a co-training version of CNB was presented too. The co-training process was a 4-step process:

1. Train CNB classifier with training dataset.
2. Classify test dataset and take messages with a confident value higher than 0.9.
3. Train CNB with training dataset and messages took in step 2.
4. Classify messages in test dataset.

For this work, experiments have been presented for the following tasks:

- **Task 1: Sentiment Analysis at global level:** Consists on performing an automatic sentiment analysis to determine the global polarity of each message in the test set of the General corpus. This task has been tested in two modes. One model with 5 levels (P+,P,NEU,N,N+) and NONE. Appart from this, the systems will be checked with a 3 levels model too (POSITIVE, NEGATIVE, NEUTRAL).
- **Task 2: Topic classification:** The technological challenge of this task is to build a classifier to automatically identify the topic of each message in the test set of the General corpus.

The experiments are:

- **UNED-JRM-task1-run1:** Co-training CNB classifier for sentiment analysis.
- **UNED-JRM-task2-run2:** CNB based classifier for sentiment analysis.
- **UNED-JRM-task2-run1:** Co-training CNB classifier for sentiment analysis.
- **UNED-JRM-task1-run2:** CNB based classifier for sentiment analysis.

In fact, the classifiers are the same for both task, the only difference is the data to classify.

Once the official results where published, my experiments obtanined the results shown in Table 3. As we see, the results obtained are not very encouraging, but we can see a silver lining with the experiment related to *run2*, wich obtains similar precisions for task 1, in both 5 and 3 levels, and task2.

Results related to *run2* are results for CNB classifier, while results for *run1* are those related to co-training classifier. Probably, the reason of these poor results with CNB are because the feature selection method is not good enough and maybe another method might be used instead of TF, like TF-IDF, or allow repeated terms between topics.

On the other hand, results obtained by co-training classifier are far worse than CNB, while better results where expected, because the distribution of the corpus is more suitable for semi-supervised methods. But, if we see the results for CNB they do not fit with the explained situation. Those bad results must be associated to a bad learning phase with the training dataset. If we do not have a good learning phase with training dataset, we can not expect a good classification, thus results for co-training are worse.

5 Conclusion and future work

In this work, a classification system has been developed, whose main aim is to carry out text classification and sentiment analysis without distinguishing. In some way the goal has been reached, becasuse altough the results are not the expected ones, they are similar for text classification and sentiment analysis. This was the main goal, check how different these tasks are. This is the main reason why new researches might be focused on the classification task itself, more than on the kind of data to classify. This task, would be very interesting, specially now, that new classifing task are emerging and dispersing techniques with data specific characteristics. Of course, specific data like emoticons in sentimen analysis, can be very useful, but a good classifier would be able to obtain goog results without this data.

Although the results have not been very encouraging, our experiments ranked 30th out of 36 in task 1 with 5 levels, position 21 of 46 in task 1 with 3 levels, and position 7 of 20 in task 2. This encourages to further research in this area, trying to reach new editions of TASS with better results and larger contributions.

After carrying out this work, many future works can be proposed. As mentioned before, the corpus provided by TASS could be not ideal for training, so it could be reorganized to have an equilibrated training dataset.

Alternative methods for feature selection might be studied too. The method proposed here does not seem to be good enough, so probably another one is needed. Even, an alternative classifer could be considered, using methods based on LDA or whatever.

At the same time, many classifying systems have been proposed and studied in the laboratory, but not much has been proved in real situations. In this case, a text clasification or sentimen analysis could be applied to real time stream in Twitter.

We can see that the majority of the work is focused on the English language. Here, we are interested in Spanish language, but we are much less. In this line, we can work in a multilingual text classification system, and by extension, a sentiment analysis system too. For this task we will need an annotated dataset, wich could be developed at the same time as the classification system.

Bibliografía

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, y Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Noviembre.

Kwak, Haewoon, Changhyun Lee, Hosung Park, y Sue Moon. 2010. What is twitter, a social network or a news media? En

Experiment number	**1**	**2**	**3**	**4**	**5**	**6**
Stemming	x		x	x	x	
Process URLs	x	x		x	x	
Hashtags	x	x	x		x	
Menciones	x	x	x	x		
Features	3228	2396	3228	3228	3228	2396
Classifiers (Precision)	**Cross-validation**					
Complement Naïve Bayes	53,9825	55,4509	53,4285	53,9687	54,038	54,6613
Support Vector Machines	37,6091	37,6091	37,6091	37,6091	37,6091	37,6091

Table 1: Configuración de los experimentos

Topic	**Messages**
Politics	3119
Other	2337
Entertainment	1677
Economy	942
Music	566
Soccer	252
Films	245
Technology	217
Sports	113
Literature	99
Total	**9567**

Table 2: Topic distribution of training dataset

Experiment	**Task**	**Precision**
task1-run2	Task 1 - 5 levels	0.393
task1-run1	Task 1 - 5 levels	0.126
task1-run2	Task 1 - 3 levels	0.496
task1-run1	Task 1 - 3 levels	0.230
task2-run2	Task 2	0.479
task2-run1	Task 2	0.158

Table 3: Official results obtained in TASS 2013

Proceedings of the 19th international conference on World wide web, WWW '10, páginas 591–600, New York, NY, USA. ACM.

Laboreiro, Gustavo, Luís Sarmento, Jorge Teixeira, y Eugénio Oliveira. 2010. Tokenizing micro-blogging messages using a text classification approach. En *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, páginas 81–88, New York, NY, USA. ACM.

Liu, Bing. 2006. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Martínez-Cámara, Miguel Ángel García Cumbreras;M. Teresa Martín Valdivia;L. Alfonso Ureña López;Eugenio. 2012. Sinai en tass 2012. *Procesamiento del Lenguaje Natural*, 50(0).

Platt, John C. 1999. Advances in kernel methods. MIT Press, Cambridge, MA, USA, capítulo Fast training of support vector machines using sequential minimal optimization, páginas 185–208.

Rennie, Jason D. M., Lawrence Shih, Jaime Teevan, y David R. Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. En *In Proceedings of the Twentieth International Conference on Machine Learning*, páginas 616–623.

Ribeiro, Fernando Batista;Ricardo. 2012. Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Procesamiento del Lenguaje Natural*, 50(0).

Santos, Antonio Fernández Anta;Luis Núnez Chiroque;Philippe Morere;Agustín. 2012. Techniques for sentiment analysis and topic detection of spanish tweets: Preliminary report.

Sriram, Bharath, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, y Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. En *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, páginas 841–842, New York, NY, USA. ACM.

Tamara Martín-Wanton, Jorge Carrillo. 2012. Uned en tass 2012: Sistema para la clasificación de la polaridad y seguimiento de temas.

Villena-Román, Sara Lana-Serrano;Eugenio Martínez-Cámara;José Carlos González-Cristóbal;Julio. 2012. Tass - workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50(0).

Zhao, Wayne Xin, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, y Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. En *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, páginas 338–349, Berlin, Heidelberg. Springer-Verlag.