LSA based approach to TASS 2013*

LSA aplicado a TASS 2013

A. Montejo-Ráez
University of Jaén
23071 - Jaén (Spain)
amontejo@ujaen.es

M. C. Díaz-Galiano
University of Jaén
23071 - Jaén (Spain)
mcdiaz@ujaen.es

M. García-Vega
University of Jaén
23071 - Jaén (Spain)
mgarcia@ujaen.es

Resumen: Este trabajo describe la participación del equipo CESA del grupo SINAI en el workshop TASS 2013 organizado en el marco del congreso de la SEPLN de ese mismo año. Nuestro sistema propone una solución basada en recuperación de información usando el método de Análisis de Semántica Latente (siglas LSA en inglés). Los resultados no son alentadores, pero el método abre la puerta a una nueva forma de utilizar la información disponible en la web social para crear recursos que pueden ser usados en clasificación de la polaridad.
Palabras clave: Análisis de Emociones, TASS 2013, LSA, Web Social

Abstract: This work describes the participation of the CESA team of the SINAI research group in the TASS 2013 workshop, organized as part of the SEPLN congress in 2013. Our system proposes a solution based on Information Retrieval, by applying Latent Semantic Analysis. Results are not very promising compared to other competitors, but the method opens a new approach in the use of social web publications as resource for sentiment polarity classification.
Keywords: Sentiment Analysis, TASS 2013, LSA, Social Web

1 Introduction

TASS stands for Sentiment Analysis Workshop (from its acronym in Spanish). This year, additional tasks have been included in the challenge compared to previous year (Villena-Román et al., 2012). All information about the workshop can be found at the official web site[1].

Our approach, based on LSA as described in next section, takes its train data from the continuous stream of posts from Twitter, capturing those that are likely to include affective expressions and generating a corpus of "feelings" that are labeled acording to their polarity. No training data from controlled corpora have been used, as we belive that trained models suffer from domain-related limitations. In TASS 2013 competition we evaluated how good our approach is compared to state-of-the-art machine learning based methods.

2 System architecture

Despite the fact that intensive feature extraction, transformation, feature generation and selection (Siqueira and Barros, 2010; Davidov, Tsur, and Rappoport, 2010) is considered in sentiment analysis tasks, we explore a new way to construct a polarity classification system by considering the continuous flow of posts in Twitter as main source for learning affective models. The system consists of the following components:

- A text processing component to convert tweets into vectors according to the Vector Space Model (Salton, Wong, and Yang, 1975) (TF.IDF weighting) after a normalization process
- An automatic extraction system to generate a term-feeling matrix from affective Twitter posts
- A linear algebra operation to reduce the term-feeling matrix generated
- A polarity calculation to propose a polarity value for tweets in the testing set

* This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), TEXT-COOL 2.0 project (TIN2009-13391-C04-02) and ATTOS project (TIN2012-38536-C03-0) from the Spanish Government. Also, this paper is partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607).

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

2.1 Tweet normalization and vector generation

Every collected tweet (see next section) and every tweet from the test set has to be converted into a vector of weighted terms. The TF.IDF weighting scheme is followed, but terms are extracted by following these steps:

1. Text are tokenized preserving hashtags, urls, mentions and other Twitter related items
2. Abbreviations (*que* insted of *q* for example), laughings (*ja* instead of *aajajaja* among others) and emoticons (e.g. *_POSITIVE_* instead of *:-)*) are processed, so certain expressions as replaced by a canonical form according to a translation dictionary.
3. A speller with additional lexicons with Spanish locations, interjections and named entities parses the text, correcting some informal words (e.g. *cuando* instead of *kuando*)

2.2 Term-feeling matrix generation

This approach is similar to that used in (Montejo-Ráez, 2012), where the WeFeelFine[2] data is used to generate a corpus of sentiments, and a the Lucene engine[3] is used to generate a ranked list of sentiments from a given text as query.

In our proposed system, a Spanish corpus has been generated by filtering those tweets with the expression "*Me siento **" as pattern, storing the adjective in the place of the asterisk as the feeling representing that text. In this way, we have collected thousands of affective tweets. This corpus was tested on sentiment analysis tasks using also Lucene search engine as described in (Montejo-Ráez et al., 2013).

The tweets were retrieved during 35 days, between December 2012 and January 2013, collecting a total of 1,863,758 tweets. Figure 1 illustrates the approach carried out to generate the *MeSiento* corpus, which is explained in detail below.

The *MeSiento* corpus was analyzed while it was being generated. From the total of 1,863,758 tweets collected, 1,516,184 tweets were not a *retweet* (RT). The number of sentiment words selected was 201, of which 84 were considered as positive and 117 as negative. The total number of different unified forms was 344, while the total number of different words (*positive* + *negative* + *neutral*) was 538. Figure 2 shows the number of tweets retrieved per hour (in Spain time zone).

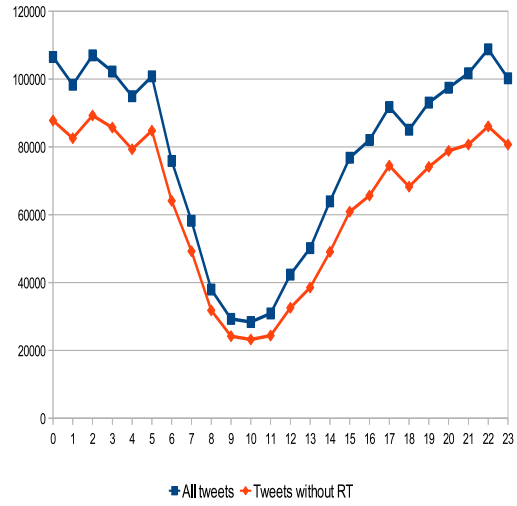

Figure 2: Number of tweets retrieved per hours

2.3 Term-feeling matrix reduction

In order to perform Latent Semantic Indexing(Deerwester et al., 1990), term-feeling matrix has to be generated. To this end, feelings are represented in the columns and terms in the rows. Singular Value Decomposition (SVD) is calculated by using linear algebra functions in the SciPy[4] library. When reducing the matrix of singular values, lower rank values are taken in such a way that 90% of the variance is preserved.

2.4 Final polarity value calculation

Once the term-feeling matrix is reduced by SVD, new tweet vectors can be multiplied by this matrix obtaining a ranked list of feelings by cosine distance, as expressed in Equation 1:

$$p(\vec{t}) = \frac{1}{|F|} \sum_{\vec{f} \in F} \frac{\vec{t} \cdot \vec{f}}{|\vec{t}| \cdot |\vec{f}|} \cdot l_f \qquad (1)$$

[2] http://www.wefeelfine.org
[3] http://lucene.apache.org/
[4] http://docs.scipy.org

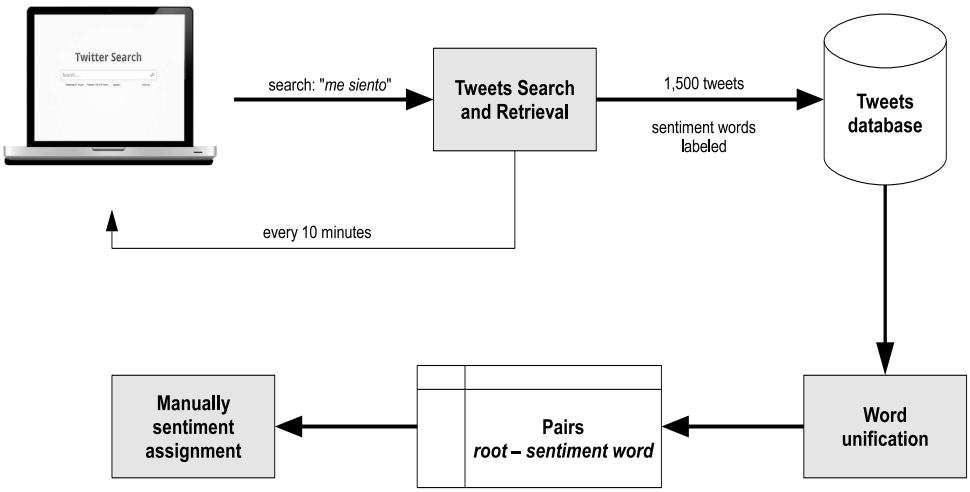

Figure 1: Steps followed to generate the *MeSiento* corpus

where $\vec{t}$ is the tweet vector, F is the set of collected feelings vectors (the columns of the reduced matrix), $\vec{f}$ is the vector of the feeling f and l_f is the polarity label (manually assigned) of feeling f (which can be 0, 1 or -1).

3 Tasks and results

We describe now how the different tasks are faced by applying the system described.

3.1 Task 1: Sentiment Analysis at global level

We have no distinguished beyond three polarity levels: Positive, Neutral or Negative, therefore, our submissions are the same for the 5-level and 3-level classification tasks in tasks 1 and 3. We have submitted two runs: one with tweet normalization (as described below) and one without normalization, generating the term-feeling matrix directly from the terms in tweets.

Results obtained are shown in Table 1. These values are somewhat discouraging, as 0.686 in F1 score was obtained as best result among all participants. This moves us to further study our approach as in other experiments with a similar system results where significantly better.

Precision	Recall	F1
0.389	0.388	0.388
0.388	0.387	0.387

Table 1: 3-level task 1 results

3.2 Task 2: Topic classification

For this task, the solution adopted is straighforward: the training set is used to generate a term-topic matrix, so for each tweet in the test set, the topic with highest similarity (minimal distance) is returned.

This simple approach shows its bad performance (due to the reduced size of the training set) in the results obtained (see Table 2). It is clear that, here, further information sources should have been used to generated better models.

Normalization	Precision	Recall	F1
Yes	0.161	0.159	0.160
No	0.161	0.159	0.160

Table 2: Task 2 results

3.3 Task 3: Sentiment Analysis at entity level

We have not performed further analysis at entity level, so the polarity assigned here is the same as for task 1. According to official results, it seems clear that trying to produce a more refined polarity at entity level is a very hard task, as performance values are very similar for all the submissions and participants.

Normalization	Precision	Recall	F1
Yes	0.384	0.384	0.384
No	0.376	0.372	0.374

Table 3: Task 3 results

3.4 Task 4: Political tendency identification

In this case, we have adopted the following solution:

- The average polarity of each user per political party is computed
- Each party has been categorized manually into *right*, *left* or *center* tendence (see Table 4)
- The tendence is assigned according to the most positive tendence of the user according to the average values of polarity of user-party tweets.

For example, user *arsenioescolar* has 24 tweets about PSOE and 64 about PP. The tweets about PSOE have a polarity count of -12, whether the PP related tweets have a count of -30. As -12 is "more positive" than -30, then the user is selected to be closer to PSOE and we considered that party to the *left* tendence.

In some cases, the final count is 0 for the "most positive" party. We have submitted 4 runs, with normalization and no normalization (as in previous submissions) and also depending whether the 0 count produces an "undefined" label or not.

Party	**Tendence**
PP	right
PSOE	left
CiU	right
IU	left
UPyD	center

Table 4: Political tendence assigned to main political parties

Undefined	Norm.	Precision	Recall	F1
No	No	0.583	0.399	0.474
No	Yes	0.570	0.386	0.460
Yes	Yes	0.467	0.316	0.377
Yes	No	0.444	0.304	0.361

Table 5: Task 3 results

4 Conclusions

We have followed a method with no intensive machine learning processing that has been trained out from the training set given by TASS organizers. Instead of existing sentiment analysis related resources, we have performed an extraction of affective tweets according to the expression "*me siento **".* From this, a term-feeling matrix is constructed and reduced by performing LSA. Applying a product of tweet norm to this matrix generates a list of ranked feelings. This ranked list allows us to compute a final polarity value thanks to a manually generated polarity labeling of the 200 feelings considered.

The results obtained are not that of state-of-the-art methods, but seem consistent with other participants. Anyhow, it encourages us to further study these results and improve our method.

References

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deerwester, Scott C., Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Montejo-Ráez, Arturo. 2012. Wefeelfine as resource for unsupervised polarity classification. *Procesamiento del Lenguaje Natural*, 50(0).

Montejo-Ráez, Arturo, Manuel Carlos Díaz-Galiano, José Manuel Perea-Ortega, and Luis Alfonso Ureña-López. 2013. Spanish knowledge base generation for polarity classification from masses. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 571–578. International World Wide Web Conferences Steering Committee.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Siqueira, Henrique and Flavia Barros. 2010. A feature extraction process for sentiment analysis of opinions on services. In *Proceedings of International Workshop on Web and Text Intelligence*.

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