TASS 2017: Workshop on Semantic Analysis at SEPLN, septiembre 2017, págs. 43-49

OEG at TASS 2017: Spanish Sentiment Analysis of tweets at document level

OEG en TASS 2017: Análisis de Sentimientos de tweets en español a nivel de documento

María Navas-Loro
Universidad Politécnica de Madrid
Ontology Engineering Group
Campus de Montegancedo
28660 Boadilla del Monte, Madrid
mnavas@fi.upm.es

Víctor Rodríguez-Doncel
Universidad Politécnica de Madrid
Ontology Engineering Group
Campus de Montegancedo
28660 Boadilla del Monte, Madrid
vrodriguez@fi.upm.es

Abstract: This paper describes the 'oeg' submission to task 1 of the TASS 2017 workshop, focusing on Sentiment Analysis at tweet level. Different parameters and systems were tested in each one of the three corpora released for the task, including different Machine Learning algorithms and morphosyntactic analyses for negation detection, along with the use of lexicons and dedicated preprocessing techniques for detecting and correcting frequent errors and expressions in tweets. The obtained results offer a basis for the design of future strategies for systems to tackle Sentiment Analysis in Twitter.
Keywords: Machine Learning, Sentiment Analysis, Polarity, TASS, Twitter

Resumen: Este artículo describe la participación del equipo 'oeg' en la tarea 1 del workshop TASS 2017, enfocado al análisis de sentimientos en tweets. Se evaluó el desempeño de diferentes sistemas bajo diferentes configuraciones para cada uno de los tres corpus propuestos para la tarea, incluyendo distintos algoritmos de aprendizaje automático y análisis morfosintáctico para la detección de la negación, así como el uso de lexicones y de técnicas de preprocesamiento específicas para corregir y detectar errores y expresiones frecuentes en tweets. Los resultados obtenidos sirven de base para diseñar la estrategia a seguir en futuros sistemas para el análisis de sentimientos en Twitter.
Palabras clave: Aprendizaje Automático, Análisis de Sentimientos, Polaridad, TASS, Twitter

1 Introduction

Recent boom on Sentiment Analysis, partly due to Natural Language Processing (NLP) new techniques and to the wide use of social networks in everyday life by Internet users, has derived into the creation of new resources and techniques to analyze opinions in almost every possible field. One of the evidences of this growing importance of Opinion Mining is its use by brands in order to discover customer's opinions. These opinions can be found in posts in the social media, being possible to some extent to automatically evaluate their polarity and the impact of marketing campaigns. According to Nielsen (Nielsen, 2012), up to 70% of users take into account the product experience published by other users, being this analysis therefore extremely valuable for companies.

The OEG, together with Havas Media, has participated in the LPS-BIGGER project[1], where software components have been developed for the categorization of brand-related messages into four categories framed in marketing analysis, being one of them a sentiment analysis task. This software is capable of classifying Twitter messages in Spanish and English into one or more of eight pre-defined emotions (love-hate, satisfaction-dissatisfaction, trust-fear, happiness-sadness). An adaptation of this infrastructure has been used to detect polarity in the Spanish messages proposed by Task 1 of the TASS 2017 workshop. The TASS workshop (Martínez-Cámara et al., 2017) has

[1] http://www.cienlpsbigger.es/

ISSN 1613-0073

become one of the first efforts to cover Sentiment Analysis in Spanish, challenging since 2012 both researchers and industry to analyze different annotated or tagged corpora. In this edition, two tasks have been proposed, but the OEG participation just covers the first one, dealing with Sentiment Analysis at tweet level.

The reminder will be as follows. Section 2 covers related works in the area, including both general approaches and proposals. Section 3 exposes our proposals in detail. Section 4 includes the results and an analysis on them, and Section 5 presents our conclusions on our participation and future lines.

2 Related Work

Several authors, such as Pang and Lee (Pang, Lee, and others, 2008) and Liu (Liu, 2010), have published comprehensive reviews of research in Sentiment Analysis, often seen as a classification problem (understood as the task of classifying a text document on a bunch of predefined categories (Liu, 2010; Kleinginna and Kleinginna, 1981)) and therefore addressed with different approaches:

- Rule-based systems (Ding and Liu, 2007; Kan, 2011). Applied both on plain texts and on POS-annotated texts, they usually rely on sentiment lexicons relating lexical units to sentiments. They usually present good precision results but require a big set of rules to get a great recall.
- Machine Learning systems (Mullen and Collier, 2004; Turney, 2002), using supervised or unsupervised techniques. They are usually trained with different kind of features, such as the words of the sentence, their lemmas, or n-grams. These systems require of a training process, but they are usually unable to capture irony and other more complex linguistic phenomena. Classical algorithms used in this approach are Naïve Bayes (Minsky, 1961) and Support Vector Machines (Cortes and Vapnik, 1995); more modern approaches include recent NLP proposals such as Word Embeddings (Mikolov et al., 2013), as happens in (Giatsoglou et al., 2017). Also different ways of handling texts, such as using Bag of Words and Bag of Lemmas, can be found along with the use of lexicons and FSS (Feature Subset Selection) techniques, since they have demonstrated to be useful for Sentiment Analysis tasks (Gamon, 2004).
- Hybrid Systems (Pang, Lee, and others, 2008; Prabowo and Thelwall, 2009), trying to avoid handicaps of each of the previous approaches.

All these kind of systems usually rely on lexicons to enrich post representation, such as SentiWordnet (Esuli and Sebastiani, 2006), the MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon (Wiebe, Wilson, and Cardie, 2005), and the Harvard General Inquirer (Stone et al., 1968) for English, associating polarity values to lexical atoms. However, the polarity of words varies from its use context and without a term disambiguation, the value of sentilexicons is limited.

Additionally, analyzing social media posts has a complex plus due to the usual presence of irony and other linguisitc phenomena of the sort (Chatzakou and Vakali, 2015), and the fact of post being often written fast and in an informal way, leading both to typos and to the inclusions of symbols, shorthands, emoticons or slung expressions that change every day following different Internet trends.

Nevertheless studies and resources on languages different from English are scarce. Some examples of works for Spanish are the adaptation performed in (Brooke, Tofiloski, and Taboada, 2009) of the system described in (Taboada et al., 2011) by translating the core lexicons and adapting other resources in various ways; also (Sidorov et al., 2012) presented an analysis of various parameter settings for most popular machine learning classifiers for the Spanish language; finally, syntactic structure of the text was used in (Vilares, Alonso, and Gómez-Rodríguez, 2013) to tackle negation, among others. A corpus in the specific field of sentiments towards brands has recently appeared (Navas-Loro et al., 2017).

3 Classification approach

The described system uses a Machine Learning approach with a Java-based standard pipeline. Some linguistic aspects such as negation and some different labeling strategies have been explored in more depth.

3.1 Labeling strategy

A binary classification system assigns or not a label to a text document, e.g. the classification of an email text as ham/spam. The labeling strategy is a straightforward operation –when a threshold is reached the message is categorised as spam, otherwise the message is classified as ham. In case of classifying the polarity of a text, the labeling strategy is somewhat more complex, as the number of possible results is at least three, positive, negative or neutral. For the TASS task 1, there is one category more: "no sentiment", which is different from neutral.

In order to classify a message as P (positive), N (negative), NEU (neutral) and NONE (no polarity), different strategies can be followed:

1. Labeling each document with one of the four categories independently and using four binary classifiers. The category is assigned to the class with the highest score among the four results. In order to have an optimal Bayesian classifier, the results for each class can be weighted by their a priori probabilities.
2. Labeling each document as either positive/non-positive and negative/non-negative. Two binary classifiers are used, and depending on the individual classifier results (P and N) the classified category is determined as follows:
   - If a tweet has similar values for P and N (this is, the distance $d = |P - N|$ is less than some threshold t_d), we can consider it as neutral (NEU).
   - If a tweet has despicable values for P and N but not extremely close to zero (this is, N and P values area between some some thresholds $t_{n_{min}} < P < t_{n_{max}}$ and $t_{n_{min}} < N < t_{n_{max}}$ with $0 < t_{n_{min}}$), we can consider it as neutral (NEU).
   - If a tweet has close-to-zero values for P and N (this is, N and P are such that $P \leq t_{n_{min}}$ and $N \leq t_{n_{min}}$), we can consider it as NONE.
3. Labeling each document with one of the three tags NEU, P and N, and then using one single classifier. The classifier score determines if the document is classified as P (high values), as N (low values) or as NEU (intermediate values). The system would never return NONE (which is only present in 14% of the documents in the InterTASS corpus). Ignoring neutral values has been reported as a bad strategy, as studied by (Koppel and Schler, 2006).
4. Using a multi-sentiment classifier, where some categories (happiness, love, satisfaction, trust) lead to choosing P, and some categories lead to choosing to N (sadness, hate, dissatisfaction, fear).
5. Using a two-stages classifier, where the first one determines the subjectivity (sentiment/no sentiment) and the second determines the polarity, as proposed by (Wilson, Wiebe, and Hoffmann, 2005).

3.2 Linguistic considerations

During our processing, diverse linguistic features were taken into account.

3.2.1 Features and negation treatment

We tested two different lexical units in our systems: tokens and lemmas. In some tested systems we also considered as features words extracted from lexicons, and also the presence of negation in verbs, detected by using trees extracted by deep syntactic analysis.

Negation was tackled by detecting the presence of "NEG" constituents in the verbal groups, with the logic that if a negation is present within a verbal group, the polarity is inverted. Double negation was not considered.

3.2.2 Preprocessing

For preprocessing the tweets, often full of grammar errors and social networks expressions that are highly decisive in polarity (such as emoticons), we have developed a filter able to detect these phenomena partly, similar to the one described by Quiros et al. (Quirós, Segura-Bedmar, and Martínez, 2016). More concretely, it is able to recognize:

- Several laugh patterns (*"hahaha"*, *"ja"*...).
- URL formats (in order to delete them, since they give no information).
- Slang expressions and replacements in Spanish social networks, such as:

  - *'q', 'k', 'qu', 'ke', 'qe'* → *'que'*.
  - *'d'* → *'de'*.
  - *'tb'* → *'también'*.
  - *'lol'* → *'ja'*.
  - *'xq', 'pq', 'porq'* → *' porque'*.
- Typos related to repeated letters (*'LOOOOOL'* → *'LOL'*)
- Suppression of numbers, as they just tend to carry polarity in concrete expressions.
- Emoticons detection and polarization, such as:
  - positive polarity : { ':-)', ';)', ':D', '<3', ':<', ':P', 'o:', '*.*' }.
  - negative polarity : { ':'(', ':S', ':$', ':_(', ':C' }.

Additionally, for some systems a stopword filter was tested. This filter based on a list of common words that carry no semantic or polarity meaning, created combining the results of algorithms (such as TF-IDF) and manual revision.

3.3 Means

3.3.1 External resources

IXA-pipes (Agerri, Bermudez, and Rigau, 2014) has been the NLP suite of choice, being the POS tagger, lemmatizer and constituent extractor the key components that were used. Weka (Frank, Hall, and Witten, 2016) has been used as the implementation of the machine learning algorithms, given its flexibility and matureness.

In order to assess the affectivity of documents, a dataset of Spanish words and their arousal (the level of activation or intensity that a stimulus elicits) and valence (how pleasant a stimulus is) was also used. The 875 words studied by Hinojosa et al. (Hinojosa et al., 2016) were lemmatized and matched against the lemmas in the document. The sum of the matched tokens' arousal was normalized and compared against a threshold to determine whether a non polarized message was NEU or NONE. Equivalently, the normalized value of the sum of the matched tokens' valence was used as an additional feature for the feature vector. Both uses of the dataset of Hinojosa proved to be effectiveless, as a low percent of the messages actually matched any word in the dictionary and results did not improve. Arguably, other larger datasets might have been used (Stadthagen-Gonzalez et al., 2017).

3.4 Algorithms used

Naïve Bayes is a generative model assuming that features are independent given a class and that calculates the probability given a class using Bayes theorem. Even when independence does not happen in Natural Language, this technique has shown to deliver good results in NLP. Multinomial version of Naïve Bayes is specifically performant when dealing with language, being lexical units (words, lemmas...) frequency the data frequently used. Also SMO (Sequential Minimal Optimization for Support Vector Machines) classifiers have been tested, but results did not improve those from Naïve Bayes.

4 Final systems and results

We present now the configurations that turned out to be the best ones among the different options exposed before:

- *laOEG* implements the second label-based strategy described in section 3.1, which performed better than any of the other approaches. Several thresholds were tested: Multinomial Naïve Bayes algorithm's threshold was set to 0.30 (also executions with values from 0.01 to 0.50 were performed), while different values were tested for internal thresholds, with values such as $t_d = 0.10$, $t_{n_{max}} = 0.15$ and $t_{n_{min}} = 0.10$. The feature vector was built simply using tokens after the pipeline described above. The Multinomial Naïve Bayes performed slightly better than the SMO classifier, being also one of its strengths its versatility (training and features can be easily managed in different manners) and its velocity, being the fastest system among the proposed; however, besides these strong aspects, it is also the simplest approach, processing just shallowly at token level and being therefore unable to detect nuances such as negation or irony.
- *victor0* Same as above, but using lemmas in the feature vector and considering negation. The presence of negation in the verbal group at different constituent levels led to the addition of extra

features (e.g. 'don't like' is handled as a single feature instead as three words or lemmas in a bag) .

- *victor2* Same as *victor0*, but also considers the presence of stopwords and using the Hinojosa dataset (Hinojosa et al., 2016) to better distinguish between NEU and NONE.
- *victor3b* Same as *victor2*, but using IBM Watson Natural Language Understanding[2] module when its output was clear (confidence level bigger than 0.75).

Numerical results of the systems proposed by OEG are exposed in Table 1 (for the InterTASS corpus), Table 2 (for the full test General Corpus of TASS) and Table 3 (for the 1k General Corpus of TASS), along with the highest and the lowest results of the overall of the participants for each corpus.

System	M-P	M-R	M-F1	Acc
victor2	0.400	0.389	0.395	0.451
victor0	0.388	0.378	0.383	0.433
laOEG	0.383	0.370	0.377	0.505
Max. result	0.497	0.490	0.493	0.607
Min. result	0.291	0.322	0.306	0.479

Table 1: In the first part of the table, Macro-Precision, Macro-Recall, Macro-F-Measure and Accuracy results for OEG systems against the InterTASS Test corpus. The second part includes the first and the last classified systems of the global ranking of participants for the same corpus.

System	M-P	M-R	M-F1	Acc
victor2	0.395	0.384	0.389	0.496
laOEG	0.350	0.342	0.346	0.407
Max. result	0.559	0.595	0.577	0.645
Min. result	0.302	0.348	0.324	0.434

Table 2: In the first part of the table, Macro-Precision, Macro-Recall, Macro-F-Measure and Accuracy results for OEG systems against the General Corpus of TASS (3 levels, full test corpus). The second part includes the first and the last classified systems of the global ranking of participants for the same corpus.

System	M-P	M-R	M-F1	Acc
victor3b	0.402	0.337	0.367	0.386
victor2	0.361	0.370	0.366	0.412
laOEG	0.348	0.345	0.346	0.448
Max. result	0.559	0.595	0.577	0.645
Min. result	0.302	0.348	0.324	0.434

Table 3: In the first part of the table, Macro-Precision, Macro-Recall, Macro-F-Measure and Accuracy results for OEG systems against the General Corpus of TASS (3 levels, 1k corpus). The second part includes the first and the last classified systems of the global ranking of participants for the same corpus.

5 Conclusions

The mere adaptation of a differently purposed classifier does not yield optimal results for the TASS challenge. However, our results have proved to be one of the most stable among the three corpora for testing, since we obtained similar results with each of the systems in all of them, fact that is not common to other participant's proposals. Groups being the first classified in a corpus can be in the last half of the ranking in another one. We acknowledge that corpora from previous TASS editions should have been used, and that additional machine learning approaches, training and semantic resources and sentilexicons are needed.

External out-of-the-box software did not prove to work any better. IBM Waton's Natural Language Understanding was tested, because no train is needed, Spanish language is covered and emotion and sentiment analysis functionality is ready to be used. However, its results did not prove any better than the system described in this paper.

The distinction between NEU and NON is a very specific feature of this challenge that justifies specific research on the strategies presented in Section 3.1. Also, in future TASS editions, special focus will be given to word sense disambiguation, introducing concepts as tokens rather than simple words or n-grams; and we will extensively use the broader sentilexicons newly appeared.

[2]https://www.ibm.com/watson/developercloud/doc/natural-language-understanding/

Agradecimientos

This work has been partially supported by LPS-BIGGER (IDI-20141259, Ministerio de Economía y Competitividad), a research assistant grant by the Consejería de Educación, Juventud y Deporte de la Comunidad de Madrid partially founded by the European Social Fund (PEJ16/TIC/AI-1984) and a Juan de la Cierva contract.

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *LREC*, volume 2014, pages 3823–3828.

Brooke, J., M. Tofiloski, and M. Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *RANLP 2009*.

Chatzakou, D. and A. Vakali. 2015. Harvesting opinions and emotions from social media textual resources. *IEEE Internet Computing*, 19(4):46–50.

Cortes, C. and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Ding, X. and B. Liu. 2007. The utility of linguistic rules in opinion mining. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812. ACM.

Esuli, A. and F. Sebastiani. 2006. SENTIWORDNET: A high-coverage lexical resource for opinion mining. In *Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)*.

Frank, E., M. A. Hall, and I. H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical machine learning tools and techniques"*. Morgan Kaufmann.

Gamon, M. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics.

Giatsoglou, M., M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Hinojosa, J. A., N. Martínez-García, C. Villalba-García, U. Fernández-Folgueiras, A. Sánchez-Carmona, M. A. Pozo, and P. Montoro. 2016. Affective norms of 875 spanish words for five discrete emotional categories and two emotional dimensions. *Behavior research methods*, 48(1):272–284.

Kan, D. 2011. Rule-based approach to sentiment analysis at ROMIP 2011. In *Sentiment analysis track at ROMIP2011*.

Kleinginna, P. R. and A. M. Kleinginna. 1981. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379.

Koppel, M. and J. Schler. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109.

Liu, B. 2010. Sentiment analysis and subjectivity. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing*. Chapman and Hall/CRC, pages 627–666.

Martínez-Cámara, E., M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, and J. Villena-Román. 2017. Overview of tass 2017. In J. Villena Román, M. A. García Cumbreras, D. G. M. C. Martínez-Cámara, Eugenio, and M. García Vega, editors, *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minsky, M. 1961. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

Mullen, T. and N. Collier. 2004. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of EMNLP 2004*, pages 412–418.

Navas-Loro, M., V. Rodríguez-Doncel, I. Santana-Perez, and A. Sánchez. 2017. Spanish Corpus for Sentiment Analysis towards Brands. In *Proc. of the 19th Int. Conf. on Speech and Computer (SPECOM)*, pages 680–689.

Nielsen. 2012. The social media report.

Pang, B., L. Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Prabowo, R. and M. Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

Quirós, A., I. Segura-Bedmar, and P. Martínez. 2016. Labda at the 2016 tass challenge task: Using word embeddings for the sentiment analysis task. In *TASS@ SEPLN*, pages 29–33.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon. 2012. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence*, pages 1–14. Springer.

Stadthagen-Gonzalez, H., C. Imbault, M. A. P. Sánchez, and M. Brysbaert. 2017. Norms of valence and arousal for 14,031 spanish words. *Behavior research methods*, 49(1):111–123.

Stone, P., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1968. The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1):113–116.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Vilares, D., M. A. Alonso, and C. Gómez-Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento del Lenguaje Natural*, 50:13–20.

Wiebe, J., T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.