An Approach to Subjectivity Detection on Twitter Using the Structured Information

Juan Sixto(✉), Aitor Almeida, and Diego López-de-Ipiña

DeustoTech-Deusto Institute of Technology, Universidad de Deusto,
Avenida de las Universidades 24, 48007 Bilbao, Spain
{jsixto,aitor.almeida,dipina}@deusto.es

Abstract. In this paper, we propose an approach to the subjectivity detection on Twitter micro texts that explores the uses of the structured information of the social network framework. The sentiment analysis on Twitter has been usually performed through the automatic processing of the texts. However, the established limit of 140 characters and the particular characteristics of the texts reduce drastically the accuracy of Natural Language Processing (NLP) techniques. Under these circumstances, it becomes necessary to study new data sources that allow us to extract new useful knowledge to represent and classify the texts. The structured information, also called meta-information or meta-data, provide us with alternative features of the texts that can improve the classification tasks. In this study we have analysed the use of features extracted from the structured information in the subjectivity detection task, as a first step of the polarity detection task, and their integration with classical features.

Keywords: Twitter · Text categorization · Data mining for social networks · Subjectivity detection · Social networks

1 Introduction

Since the Twitter social network was created in 2006, it has experienced a substantial growth, having more than 100 million of daily active users and 500 million tweets every day [24] nowadays. Currently Twitter is one of the largest textual data sources used in the data mining and knowledge extraction fields of research. As a part of these fields, sentiment analysis is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [15]. Several research groups have used sentiment analysis techniques [17] over the Twitter micro-texts with an acceptable grade of success. However, the particular characteristics of Twitter (Hashtags, user references, inclusion of URLs, maximum of 140 characters) generate loosely formated texts that are difficult to analyse. Addressing this challenge requires an adaptation of the classical techniques and tools to Twitter's unique requirements, that often results in a relevant decrease of their performance.

N.T. Nguyen et al. (Eds.): ICCCI 2016, Part I, LNAI 9875, pp. 121–130, 2016.
DOI: 10.1007/978-3-319-45243-2_11

There are two possible approaches to this problem. The first one is improving the quality of the texts, in order to facilitate their automatic process. This text normalization task deals with several problems like the use of slang, word shortening, letter omissions and bad spelling [23]. The application of these techniques cleans the texts and improves the performance of the lexical analysis over them. The other approach is to improve the sentiment analysis process using the structured information [5] in addition to the tweeted text. Several researches have studied [2,20] how the external information can improve the sentiment analysis task. The obtained results show that the external information is a reliable source of knowledge about sentiment and opinion of texts.

In this paper we expand the knowledge about the structured information used in opinion mining field, and its incorporation to the text analysis classical techniques. We study their application to the polarity detection in Spanish language and specifically in the subjective detection task. The rest of the paper is organized as follows: In Sect. 2, the context of this work is presented. In Sect. 3, a Structured and Unstructured information review is presented. Section 4 covers the experimental procedures, and conclusions are introduced in Sect. 5.

2 Related Work

Our study is focused on two aspects of opinion mining: the application of the contextual information and the Spanish polarity classification. In this section, we will review the papers which our work is based on.

2.1 Contextual Applications in Sentiment Analysis

The primary objective of the Sentiment Analysis field is the automatic retrieval of subjectivity and opinion polarity. However, determining their scope is a very complex task and their areas of application are extensive. There are several surveys that summarize the main applications and the most common techniques in sentiment analysis [1,16,21].

In the field of the application of contextual information, there are several researchers who use the additional information available in social networks in the classification tasks. In 2011, Pennacchiotti and Popescu [22] presented a generic model for user classification in social media that combines linguistic features and explicit social network features. They also emphasize one of the main problems of contextual information, the difficulty of collecting the social network features of a dataset. Mislove et al. [19] analysed data on a set of Twitter users in order to compare them with the U.S. population. To this end, they developed several techniques to enrich the information available of each user, detecting the gender, the ethnicity and geographic distribution of the users. This was one of the first studies that addressed the idea of the sampling bias and the study of the dataset population as an approach to improve predictions or measurements. Bermingham and Smeaton [4] modeled the political sentiment in order to predict electoral results in Twitter, including sociolinguistic features and unconventional punctuation. In the psychiatry sphere, De Choudhury et al. [7] developed a SVM classifier that can predict the likelihood of an individual to be depressed using Twitter. This work demonstrates the potential of the social networks as a tool for measuring and predicting emotional states of the users and gives new insights about the feature measures. Some of these features, used in their research, are the diurnal trends of the users, the volume of replies and the ego networks. Jiang et al. [13] present a target-depend sentiment classifier using the relations between tweets.

2.2 Spanish Polarity Classification

During the last years, research groups have published a large amount of approaches and methods in the sentiment analysis sphere, and have generated lexicons and polarity dictionaries that facilitate the tasks. Nevertheless, these tools are language dependent. Usually these are generated in one language and, at times, are translated to some other languages. This, combined with the difficulty of establish standard linguistic rules between languages, causes a performance decrease when adapting the tools to other languages [10].

A lot of papers have been published on the field of sentiment analysis in social media, specifically focused on the Spanish language. Vilares and Alonso [1], reviewed a large quantity of bibliographic references in the Spanish scope. Also the TASS workshop [26], a satellite event of the SEPLN Conference[1], presents a huge amount of algorithms and techniques based on opinion extraction in Twitter.

3 Structured and Unstructured Information

Twitter contains a large amount of information about each tweet in addition to the tweeted text. Hashtags, retweets, replies, mentions, followers and many other relations bring us a considerable volume of information about the user network and all its components. This can be a knowledge source about users and their opinions, as we have seen in previous researches, and bring an improvement to the sentiment analysis tasks. Assuming that the use of structured information in sentiment analysis tasks has been proved, our aim is to check their efficiency in the subdomain of automatic sentiment analysis at global level. This subdomain consist of performing an automatic sentiment analysis to determine the global polarity of each message about any topic, without any previous topic discrimination. According to our research proposal, we pretend to study the new possibilities of the structured information in the global level of sentiment analysis, adapting features used in concrete domains as politics [5] or psychiatry [10], and other features not used yet.

Currently, does not exist a unequivocal terminology to refer to the contextual information of the tweets. Barbosa and Feng [2] name them as

[1] http://www.sepln.org/.

"Tweet Syntax Features" and Liu [14] refears to them as "Twitter specific clues". In this paper we use the terminology of Structured and Unstructured information, described in Cotelo et al. [5]. In order to achieve the proposed task, it's necessary a full understanding of the Twitter structure and of how their components (Users, Texts, Communities) relate among them. The Fig. 1 represents the most frequent components of the social network and their relations, considered relevants to this work. Also, as part of the study of the data, the structural information of the social network has been classified according to its origin, that is, the component where the information origins.

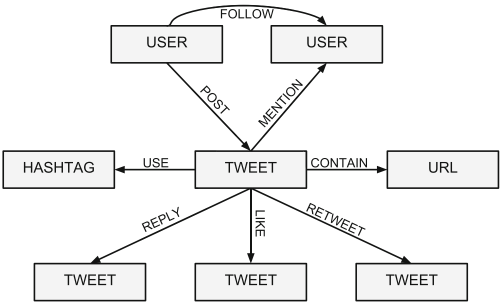

Fig. 1. Twitter structure representation.

As part of our research we have classified the available tweet data in four categories:

Text attributes. This category comprehends the attributes that appear in the text, but that does not depend on the words representation. These attributes emphasize on how the tweets are written, including the characteristic properties of Twitter that are part of the text, but not encompass semantic or sintactic analysis. Technically, this category could be considered as a subcategory of Tweet attributes, as long as the text is a part of each tweet, even so we consider this features clearly difference from the Tweet attributes. Examples include: *hathtags, links, emoticons, punctuation, retweet, used language.*

Tweet attributes. This category comprehends all the characteristics that define a only tweet but not are part of the text. These usually refer to the living process of the tweet within the network, the relations of the users with them or the way it had been posted. Examples include: *quantity of retweets, quantity of favourites, creation date/time, the application that sent the tweet, associated place.*

User attributes. This category comprehends the attributes relative to the authors of the tweets. These attributes represent several aspects of the users that may be relevant to understand the context of a tweet. Occasionally, these aspects are compiled to generate úser profileśthat simplify the user representation in the system. Examples include: *location, political affiliation, post habits.*

Topographic attributes. This category comprehends all the measures about the network topology. The topographic attributes often require some complex calculations and help us to know the role of a user or tweet in the network. Examples include: *Modularity class of user, In-degree and Out-degree of users, Network Communities of the mentioned users.*

4 Experiments

Our purpose is to predict the subjectivity of tweets using the structured components. The main characteristics of the subjectivity classification task are reviewed below. In order to detail the realized experiments, we also describe the corpus chosen, the features studied and the classification algorithms.

4.1 Subjectivity Detection

Liu [14] defines the subjectivity classification as follows: *"Subjectivity classification classifies sentences into two classes: subjective and objective. An objective sentence expresses some factual information, while a subjective sentence usually gives personal views and opinions."*. The subjectivity detection problem has been studied for several years in different areas, especially the approaches based on supervised learning. Since the beginning of the sentiment classification researches, subjectivity has also been explored as part of the global sentiment classification area. The sentiment classification can be expressed as a classification problem if three or more classes: Positive opinion, Negative opinion and no opinion, although these three classes are expanded in several cases. Furthermore, the problem can be divided into two classification subproblems; the opinion detection task first and the distribution between positive and negative opinions later on.

In this paper we address this first problem that is usually named Subjectivity Detection problem. In the Twitter research area, several authors have worked on the sentiment classification in which the subjectivity detection plays an important role. Barbosa and Feng [2] use some twitter features to implement a subjectivity classification. Davidov et al. [6] propose a sentiment classifier that uses punctuation-based features in posted texts. Due to their short length, each tweet is considered as a single sentence and accordingly, each tweet has only a single sentiment polarity.

4.2 Selected Corpus

The scope of our study is focused on the detection of subjectivity in Tweet texts in Spanish language. At present, only a few number of datasets that satisfy these conditions can be found in the state of art. Only the multi-language dataset presented in Volkova et al. [27] and the TASS dataset [26] includes texts with no sentiment in Spanish language.

Finally we decided to use the datasets of the sentiment analysis task at TASS'15[2] workshop [26]. This is an evaluation workshop for sentiment analysis focused on Spanish language, organized as a satellite event of the annual conference of the Spanish Society for Natural Language Processing (SEPLN). This paper is focused on the first task of the workshop, that consists in performing an automatic sentiment analysis to determine the global polarity of each message in the provided corpus. Tweets are divided into six different polarity labels: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE). Tweets are also divided into two sets, the training dataset with 7.219 (11 %) items and the test dataset with 60.798 (89 %) items. Additionally, the task includes a 1.000 items dataset, a balanced and hand-labeled subset of the test dataset, that we use as evaluation of the performance of our systems.

The dataset contains a 20.54 % (Train) and 12,30 % (Test) of tweets tagged as NONE, that are considered as True results by our classifier, whereas the rest of labels are considered as False results. As our aim is the detection of objective and subjective texts as a first step of the polarity detection task, the use of a balanced or binary corpus was discarded. The significance of obtaining not only a high accuracy but also a high F1 measures have been taken into account and are explained in the Experimental Results section.

4.3 Features

Several features have been used to address the subjectivity classification problem [14], the vast majority based in the text: bag-of-words, vectorial representations of words, n-grams, etc. In the structural features area, some authors have studied the use of several features in the subjectivity classification of English tweets. Barbosa and Feng [2] exploit the use of retweets, hashtags, replies, links, punctuation, emoticons and the number of upper cases in a subjectivity classifier. Davidov et al. [6] use the length of the words and the punctuation signs as features.

Based on the approach to structured and unstructured information presented in by Cotelo et al. [5], we implement a single sentiment classifier for subjectivity detection, combining two classifiers, each one trained with a different type of information. As shown above, the corpus contains six different labels and as consequence it is not balanced in respect to the objectivity-subjectivity axis. This has been taken into account during the training and evaluation process. In order to train the classifier based on structural information, we have composed a feature list:

URL, Exclamation marks, Emoticons and Uppercase words.
According to the work of Barbosa and Feng [2], this features has been used in our work.

[2] Workshop on Sentiment Analysis at SEPLN Conference.

Uppercase Percent. In addition to the number of uppercase words, we used the percentage of uppercase characters of the total characters. This technique differs from the feature proposed by Barbosa and Feng [2], not only counting the words that starts with upper case, but counting all the characters. This ensures that the tweets with all capitals texts, tipically used for emphasis or "shouting", are taking into account.

Favorites. Twitter includes an option called "favorite", that allows the users to like individual tweets. Our study detected a relation between the average of "favorites" by tweet and their sentiment polarity.

Modularity Class. In Twitter, users may subscribe to other users tweets. This is known as "following" and establishes a directed graph of relations between users. During the conduct of our study, we proposed to extract the relations between the users (authors of the tweets) through their "following" relations and genetared their relation graph. Their modularity class revealed the existence of only three communities. A preliminary research shows that this communities are formed by associated individuals related to the left/right political parties or ideologies, and a third group of neutral celebrities. Used as feature, the modularity of each user generates a increase of accuracy and f1-measure in the classifier.

Graph Degrees. Some other attributes of the relation graph have been proposed and tested as features. In detail, the In-Degree and Out-Degree punctuations of the authors have proven to be useful to classification task.

RT. Twitter includes an option called"Retweet", that allows to share a message from another user. This boolean feature expresses if the analysed tweet is a "retweet" of an original tweet.

Ellipsis. During our study, it was noted that some objective tweets includes an ellipsis. In Twitter, ellipsis is often used to make observations about external information, like headlines, urls, or quotes from other users.

The second classifier was based on unstructural information. The selected model to represent the texts was the commonly used the bag-of-words [11]. This model represents each tweet as a matrix of token counts of its words.

4.4 Experimental Results

Considering that the dataset is unbalanced, as shown in Sect. 4.2, we decided to extract three different measures to evaluate the performance of the tested systems; Accuracy, Macro-F1 and NONE-F1. The accuracy is the proportion of true results among the total of the dataset, however, when the prior probabilities of the classes are very different, this metric can be misleading [12]. The macroaveraged F1-measure considers precision and recall, and provides information of how the system performs overall across the dataset. The so-called NONE-F1 are the micro-F1 measure of the NONE (or True) labels of the system. The F1-measure is considered as a relevant score for evaluating the accuracy of a test with a unbalanced dataset. Also we consider relevant to evaluate the specific F1-measure of the NONE labels, in order to rank the contribution of the classifier to a polarity detection task. Obtained results are reviewed in the Table 1.

The first approach to the task consists on a single classifier with the structural features described in Sect. 4.3. Multiple classifier models were tested, like LogisticRegression, Naive Bayes, and Random Forest, obtaining the best results with a GradientBoosting classifier [9], reaching a 70.8 % in Macro-F1 and a 43.0 % in NONE-F1. Then, a new test has been realized using the bag-of-words model, in order to contrast the performance of structured and unstructured approaches. The best results have been obtained with a LogisticRegression model with a balanced mode to automatically adjust weights, obtaining a 70.4 % in Macro-F1 and a 46.8 % in NONE-F1. The similarity of the results shows that both approaches have a relevance for the task, although the poor results involve that the task is complex.

We also investigated the chance of combining both approaches in order to improve the results of the classification task. To do this, we selected two different approaches; using both feature lists in a single classifier and a stacked generalization [28]. In the first case, the structural features of each tweet were added to their matrix representation, generating a new features list. This is a very simple way to merge both models and checks if both feature lists are directly complementary or need more complex techniques to improve the results. The best results has been realized with a GradientBoosting classifier and obtained a 69.2 % in Macro-F1 and a 43.5 % in NONE-F1. This technique does not improve the results, and in several cases decreases them, proving that is necessary the use of other techniques to merge the heterogeneous features effectively.

At least, we realized a stacked generalization work for combining both models. For the level-0 generalizer we use five different classifiers; Logistic Regression, GradientBoosting, Multinomial Naive Bayes, Random Forest and Calibrated with Isotonic Regression. Each of the classifiers were trained with both models, generating a ten classifiers array that formed the level-0 models. Then we used a Logistic Regression model for the level-1 classification model. We found that the use of regression models obtain best results, according to the presented by Ting and Witten in [25]. This approach obtained a 90.22 % in Macro-F1 and a 55.66 % in NONE-F1, being the best obtained results. This improvement implies that a complex technique, like the stacking, benefits from the heterogeneous features in relation to the other approaches.

Table 1. Results for subjectivity detection.

System	Accuracy	Macro F1	NONE-F1
Meta-Information	89.5 %	70.8 %	43.2 %
Bag-of-Words	79.3 %	70.4 %	46.8 %
MI+BoW	88.3 %	69.2 %	43.5 %
Stacking MI+BoW	89.8 %	90.20 %	55.65 %

5 Conclusions and Future Work

Our objective in this study was to learn about the contextual information, their uses at the subjectivity detection task and their application improving the text based models. Exist several previous approaches to the contextual data and we have attempted to adapt these knowledge to the global polarity detection task and to the spanish language. Our study has verified a hypothesis already applied in other social areas and expanded the knowledge relative to the contextual information, adding new ways to use the contextual information to the previous approaches of the state of art. Also we presented a contextual data classification for a better understanding of their nature and characteristics. We presented a first interaction of a subjectivity detection approach which uses some contextual elements to build its features. This approach overtakes the basic classifiers and achieves to combine the structured and unstructured information, establishing a method to complement the standard classification techniques. Although the accuracy and f1-measure are around 90 %, the poor values in the micro average reveal that exists an huge margin for improvement in the task. As future work, we want to connect our work with a complete polarity detection task, applying the extracted knowledge in other sentiment categories, exploring new contextual features. We want to perform a more extensive analysis to check more Twitter components and their relation with the different polarities, considering that distinct features could be related with only a particular sentiment category. Also, we seek to apply the contextual features with more complex models that include lexicons of polarity and semantic resources to really see the impact of them.

References

1. Alonso, M.A., Vilares, D.: A review on political analysis and social media. Procesamiento del Lenguaje Nat. **56**, 13–24 (2016)
2. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44 (2010)
3. Belkaroui, R., Faiz, R.: Towards events tweet contextualization using social influence model and users conversations. In: Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, p. 3. ACM (2015)
4. Bermingham, A., Smeaton, A.F.: On using Twitter to monitor political sentiment and predict election results (2011)
5. Cotelo, J.M., Cruz, F., Ortega, F.J., Troyano, J.A.: Explorando Twitter mediante la integracin de informacin estructurada y no estructurada. Procesamiento del Lenguaje Nat. **55**, 75–82 (2015)
6. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using Twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters (2010)
7. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: ICWSM, p. 2 (2013)
8. Esparza, S.G., OMahony, M.P., Smyth, B.: Mining the real-time web: a novel approach to product recommendation. Knowl. Based Syst. **29**, 3–11 (2012)

9. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
10. Han, B., Cook, P., Baldwin, T.: unimelb: Spanish text normalisation. In: Tweet-Norm@ SEPLN, pp. 32–36 (2013)
11. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
12. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data-recommendations for the use of performance metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 245–251. IEEE (2013)
13. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent Twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151–160 (2011)
14. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
15. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C.X. (eds.) Mining Text Data, pp. 415–463. Springer, New York (2012)
16. Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A., Montejo-Ráez, A.R.: Sentiment analysis in Twitter. Nat. Lang. Eng. **20**(01), 1–28 (2014)
17. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
18. Mejova, Y., Srinivasan, P., Boynton, B.: GOP primary season on Twitter: popular political sentiment in social media. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. ACM (2013)
19. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the demographics of Twitter users. ICWSM **11**, 5 (2011)
20. Monti, C., Rozza, A., Zapella, G., Zignani, M., Arvidsson, A., Colleoni, E.: Modelling political disaffection from Twitter data. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2013) (2013)
21. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval **2**(1–2), 1–135 (2008)
22. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to Twitter user classification. ICWSM **11**(1), 281–288 (2011)
23. Porta, J., Sancho, J.L.: Word normalization in Twitter using finite-state transducers. In: Tweet-Norm@ SEPLN, vol. 1086, pp. 49–53 (2013)
24. Smith, C.: DMR Twitter Statistic Report. Last modified 26 Feb 2016. http://expandedramblings.com/index.php/downloads/twitter-statistic-report/. Accessed 28 Mar 2016
25. Ting, K.M., Witten, I.H.: Issues in stacked generalization. J. Artif. Intell. Res. (JAIR) **10**, 271–289 (1999)
26. Villena-Román, J., García-Morera, J., García-Cumbreras, M.A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureã-López, L.A.: Overview of TASS 2015. In: Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN, vol. 1397. CEUR-WS.org (2015)
27. Volkova, S., Wilson, T., Yarowsky, D.: Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: EMNLP, pp. 1815–1827 (2013)
28. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)