Sentiment Extraction from Tweets: Multilingual Challenges

Nantia Makrynioti(✉) and Vasilis Vassalos

Athens University of Economics and Business, 76 Patission Street,
GR10434 Athens, Greece
{makriniotik,vassalos}@aueb.gr

Abstract. Every day users of social networks and microblogging services share their point of view about products, companies, movies and their emotions on a variety of topics. As social networks and microblogging services become more popular, the need to mine and analyze their content grows. We study the task of sentiment analysis in the well-known social network Twitter (https://twitter.com/). We present a case study on tweets written in Greek and propose an effective method that categorizes Greek tweets as positive, negative and neutral according to their sentiment. We validate our method's effectiveness on both Greek and English to check its robustness on multilingual challenges, and present the first multilingual comparative study with three pre-existing state of the art techniques for Twitter sentiment extraction on English tweets. Last but not least, we examine the importance of different preprocessing techniques in different languages. Our technique outperforms two out of the three methods we compared against and is on a par to the best of those methods, but it needs significantly less time for prediction and training.

1 Introduction

Users have integrated microblogging services and social networks in their daily routine, and tend to share through them increasingly more thoughts and experiences of their lives. As a result, platforms, such as Twitter, are a goldmine for the tasks of opinion mining and sentiment analysis, providing valuable information on topics of timeliness or not, by users of varying social, educational and demographic background.

In this paper, we examine sentiment analysis in Twitter with emphasis on tweets written in Greek and we suggest a method based on supervised learning. Sentiment analysis is defined as the task of classifying texts, in case of Twitter these correspond to tweets, into categories depending on whether they express

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

S. Madria and T. Hara (Eds.): DaWaK 2015, LNCS 9263, pp. 136–148, 2015.
DOI: 10.1007/978-3-319-22729-0_11

positive or negative emotion or whether they enclose no emotion at all. As a consequence, sentiment analysis solves two classification tasks, the identification of objective and subjective tweets and the categorization of the latter according to their polarity. Given a number of tweets, our task is to categorize them in three classes, positive, negative and neutral depending on the presence of features that indicate emotion or not, as most of the times this is consistent with the sentiment of the message [12].

Although recently many papers study the task of sentiment analysis and many approaches have been proposed, almost all of them regard English text and work for other languages is limited. Moreover, many studies do not report results from comparisons with other pre-existent methods and each technique is usually evaluated on a single dataset. Evaluation on different datasets, including data of more than one languages, is an interesting process, which cross-checks the performance of the methods among languages.

The contributions of our paper are summarized below:

1. We propose a novel method for classification of tweets into three categories, positive, negative and neutral, and we evaluate our classifier on real Greek and English tweets. Our method outperforms two of the three compared approaches while giving statistically indistinguishable results to the third but with significant less time.
2. We present a case study of sentiment analysis in the context of the Greek language, unlike English that are much more studied and understood. For this purpose we collected and manually annotated a corpus of posts in Greek from Twitter, in order to be used as training and test data.[1]
3. We present extensive evaluation results and comparisons to three existing methods developed for English on a Greek as well as an English dataset. The purpose of these experiments is to provide the first comparative study of different state of the art techniques over Greek data, and examine their generalizability to address multilingual challenges. We also examine the contribution of specific preprocessing and postprocessing steps through ablation tests that demonstrate the degree to which certain steps of the proposed method improve the accuracy of the system with regard to Greek or English.

The rest of the paper is organized as follows. Section 2 presents some representative approaches on the problem of sentiment analysis and Sect. 3 analyzes the data used for training and testing. In Sect. 4 at first we give an overview of our method and then we describe in detail every step of it. Results from the evaluation of the classifier and the comparative analysis are reported in Sect. 5. Finally, Sect. 6 concludes and presents ideas for future work.

2 Related Work

The mining and analysis of unstructured data from social networks has attracted considerable attention in recent years. Go et al. [9] dealt with sentiment analysis

[1] Data are available by emailing the authors.

in Twitter, but their work was limited to positive and negative sentiments, and does not involve the recognition of objective (neutral) tweets. The machine learning algorithms that were applied are Multinomial Naive Bayes, Support Vector Machines (SVM) [24] and Maximum Entropy, whereas unigrams, bigrams as well as the combination of these two were used as features. Maximum accuracy reached 83 % and was achieved with Maximum Entropy and both unigrams and bigrams. Pak and Paroubek [20] emphasized the preprocessing of tweets before classification and adopted bigrams, trigrams, negation and part-of-speech tags as features. They used entropy and introduced a variant of it called "salience" to select the most representative features. Their results show that bigrams outperform trigrams and salience discriminated n-grams better than entropy. The method described in [5] divides the classification of tweets into two stages. The first stage classifies subjective and objective tweets, while the second categorizes subjective tweets into positive and negative. Part-of-speech tags are used as features in this paper too. Dictionaries of subjective terms and syntax features of Twitter, such as hashtags, links, punctuation and words in capital letters, were also employed. The classifier used SVM and maximum error rate for the first stage reached 18.1 %, whereas for the second stage it reached 18.7 %.

Even though the paper by Pang et al. [21] is not about Twitter, it is a benchmark and a comparison point with all the studies mentioned above. The paper addresses the task of sentiment analysis in movie reviews. Features include unigrams, bigrams and negation. Multinomial Naive Bayes and Maximum Entropy were tested, but SVM achieved 82.9 %, which was the maximum accuracy. Finally, a very recent approach by Mohammad et al. [17], which used a variety of features, including ngrams, syntax, lexicon and negation features, achieved the highest average F-score (69.02 %) with a SVM classifier in SemEval 2013 (International Workshop on Semantic Evaluation) and the task of sentiment analysis in Twitter [18]. Our work falls into the same category with the aforementioned studies, but apart from the certain difference of experimenting on Greek data, we apply a different combination of features and preprocessing steps, followed by a novel postprocessing negation identification step, which attempts to recognize the structure of negation in text and reverse the given prediction, rather than affect the features used for classification. Moreover, we reproduce published methods and present comparisons of them on a multilingual fashion, experimenting on datasets from two languages, Greek and English. All the above approaches belong to the category of supervised learning, but many studies have also performed unsupervised sentiment analysis. Due to limited space, we do not mention them here.

As stated earlier, there is lack of studies concerning other languages than English and the task is not sufficiently examined from this perspective. The paper by Atteveldt et al. [4] presents a system for automatically determining the polarity of relations between actors, e.g. politicians and parties, and issues, such as unemployment and healthcare, in Dutch text. To determine the polarity of relations, the authors use existing techniques for sentiment analysis in English and show that these methods can be translated to Dutch. Another study that addresses the multilingual perspective of the task is presented by Boiy and Moens [6]. The authors propose a supervised method for sentiment analysis and perform experiments on English, Dutch and French blog reviews and forum texts. There is also work about sentiment analysis on Modern Standard Arabic at the sentence level [2]. Arabic is a morphologically-rich language in contrast to English and the authors propose some Arabic-specific features along with the more commonly used and language-independent ones. Another work by Abbasi et al. [1] performs sentiment analysis on hate/extremist group forum postings in English and Arabic, and evaluates a variety of syntactic and stylistic features for this purpose. A method on Chinese data is also proposed by Zhao et al. in [26]. We are aware of a paper regarding reputation management on Greek data [22], but it presents a commercial product very briefly and in the abstract, and cannot be reproduced. Thus, our method not only is described extensively and in detail, but is also compared with other methods in the literature.

Finally, with regard to papers that compare methods and systems of sentiment analysis, such as [10] and [3], we take a step further and present comparisons in more than one languages.

3 Data

In this section we describe the datasets that are used for training and testing. Details about the size and contents of each dataset are given by Table 1. The Greek training data were collected between August 2012 and January 2015. Part of positive and negative tweets are based on subjective terms and around 20 % of neutral tweets were gathered from accounts of newspapers and news sites. The rest were streamed randomly. Respectively, Greek test set consists of random tweets posted between October 2013 and January 2015. We used Twitter Search and Streaming API[2] for the collection. Both training set and test set were labeled by three annotators. The calculated Fleiss' kappa [7] for the training set is 0.83, which is interpreted as almost perfect agreement, whereas for the test set is 0.691, which denotes substantial agreement. We will refer to the Greek training and test set as GR–train and GR–test.

For experiments on English we use training and test data provided by the organizers of SemEval 2013 [18] for the task of sentiment analysis in Twitter. The organizers collected tweets according to popular topics, which included

Table 1. Datasets

Dataset	Positive	Negative	Neutral	Total
GR–train	1870	2940	3190	8000
GR–test	261	249	378	888
Sem–train	3287	1601	4175	9063
Sem–test	1572	601	1640	3813

[2] https://dev.twitter.com/.

named entities previously extracted by a Twitter-tuned NER system [23], and used Mechanical Turk for annotation. We will refer to SemEval training and test set as Sem–train and Sem–test respectively.

4 Overview of Approach

The approach we adopt consists of three main steps: (1) Preprocessing of data. (2) Feature engineering. (3) Reversal of classifier's prediction for a tweet due to negation identification. The proposed method takes into account not only inflection but also word stress, both characteristics of morphologically-rich languages, and suggests a novel technique to reduce the negative effect of the combination of both in classification performance. Moreover, it treats identification of negation as a postprocessing step and attempts to capture its structure, which is a much different approach than adding a special suffix to bag-of-words features that most methods do until now. The aforementioned steps are described in detail in the following sections.

4.1 Preprocessing

Preprocessing is applied to both training and test set. The first step of preprocessing is the removal of noise from the data. Elements that do not indicate the polarity of a tweet are considered as noise. Such elements are listed below. (1) URL links. (2) Mentions of other users. (3) The abbreviation RT, which indicates that a tweet is a retweet of another one. (4) Stop words, including articles and pronouns. Stop words are extremely common words, which appear to be of little value in deciding the sentiment of a text.

Because users use plethora of emoticons/hashtags, we choose to replace positive emoticons[3] with the emoticon ":)" and negative emoticons[4] with the emoticon ":(". A number of hashtags, such as #fail and #win, are also replaced with the former two emoticons. The aim of this step is to group the emoticons/hashtags in two categories and to avoid the need of importing tweets in the training set for each one of them. In addition to the above replacements, possible repetitive vowels encountered in a word are reduced to one, whereas repetitive consonants are reduced to two.

Capitalization and removal of accent marks are the next steps. An accent mark over the vowel in the stressed syllable is used in Greek to denote where the stress goes, e.g. 'καλημέρα' (good morning). In order to avoid mistakes due to omission of stress marks and incorrect use of capital letters versus lowercase letters, we remove these marks from tweets and transform them to uppercase. Stemming is the third and last step, and is used mostly to compensate for data sparseness. Stems are generated by George Ntais' Greek stemmer [19] for Greek and by Lovins stemmer [15] via the Weka data mining software [11] for English.

[3] List of positive emoticons: :-), :), :o), :], :3, :c), :>, =], 8), =), :}, :^), <3, ^_^, ;>, (:, ;), (;, :d, :D.

[4] List of negative emoticons: >:[, :-(, :(, :-c, :c, :-<, :<, :-[, :[, :{, :'(, :/ .

The previous steps are applied to the test set too. However, the preprocessing of test set involves an additional step: part-of-speech tagging. It takes place before stemming and is an auxiliary step for the process of negation identification (Sect. 4.3), which follows classification. After the replacements, we annotate the words of each tweet with part-of-speech tags, which are not taken into account by the classifier to predict the class, but are used in patterns whose intention is to detect negation. A Greek part-of-speech tagger [13] is used for the tagging process in Greek, whereas in English the Carnegie Mellon University (CMU) Twitter Natural Language Processing (NLP) tool was selected [8].

4.2 Features

Feature engineering follows the bag-of-words representation with unigrams and term presence. Due to the limit in the number of characters that compose a tweet, a unigram is enough to denote the sentiment in most of the cases. For some unigrams there is a dependency with a particular class, while others do not give any information under any circumstances about the polarity. We decided to keep only a subset of them in order to eliminate noisy features and build a simpler model. We experimented with two methods, Information Gain and Chi Squared [14]. They both gave equally good results, so Information Gain is chosen arbitrarily for the experiments displayed in sections below.

Apart from word ngrams, lexicons of subjective terms, which contain terms with association to positive and negative sentiments, may provide various features for sentiment analysis. There are plenty of subjective lexicons in English, but we are not aware of any such lexicons in Greek. Nonetheless, we attempted to create manually two simple Greek subjective lexicons, one with positive words and one with negative words according to their prior polarity. Words were derived from random tweets, not contained in GR–train or GR–test, or translated from subjective English lexicons. The positive lexicon contains 199 words and the negative one consists of 292 words. We use two simple features, the presence of positive/negative terms of such lexicons in Greek tweets and more sophisticated features, such as those proposed in [17], for English data. In the aforementioned paper, lexicon-based features proved to be useful for the task of sentiment analysis. We present our conclusions about this kind of features in Sect. 5.

4.3 Negation Identification and Polarity Reversal

Negation identification is based on patterns of part-of-speech tags combined with negation words. We attempt to identify these patterns in each tweet and store the token that is negated. For example, the Greek word 'δεν' (not) followed by a verb and an adjective constitutes a negation pattern. If a tweet contains the phrase 'Η ταινία δεν ήταν καλή' (The movie wasn't good), the former negation pattern will be identified due to the word 'δεν', the verb 'ήταν' (both correspond to wasn't) and the adjective 'καλή' (good). Then the token 'καλή', which is the one that is negated, will be stored. Nine frequent patterns are recognized for Greek and eight for English. The detection of negation aims to reverse the prediction given by the classifier for a tweet from positive to negative or from negative to positive. If the prediction is neutral, no change is made. So following the decision of the classifier, we first check if a negated token is stored for the tweet. If yes, then we examine if this token belongs to the features that are present in the tweet. Suppose we have the aforementioned tweet for which we have kept the token ‘καλή’ as the negated token. If the unigram ‘καλή’ is one of the features and its value is 1, which indicates that this feature is present in the tweet, then the appropriate reversal of polarity will be performed. Otherwise, it will not.

4.4 Challenges of the Greek Language

The Greek language has a highly inflective nature that reduces the effectiveness of usual bag-of-words features. Greek verbs and adjectives are inflected for person, number and gender, which affects mostly the suffixes of the words. The various suffixes due to inflection increase ngram features, many of which are not contained in the training set. Hence, classification performance decreases. As a countermeasure to the inflective nature of Greek, the words of each tweet are replaced by their stems, assuming that stems are enough to denote the sentiment of a tweet in most cases.

Except from inflected verbs and adjectives, stress marks used in Greek make things even more complicated. Twitter users often forget to add these marks or they add them at the wrong syllable, creating this way a number of different versions of the same word (e.g. ‘καλημέρα’ is a different unigram from ‘καλημερα’). As stated earlier, we chose to remove accent marks in order to reduce ngram features, but in case of stemming this choice may lead to mistaken predictions. Specifically, although stemming operates positively and helps the method to generalize better on unseen data (a conclusion that is drawn from the ablation tests included in Sect. 5.4), there is a case where it operates negatively: the stem of two words is the same whereas their polarities are different. For example the Greek words ‘συμφωνώ’ (agree) and ‘σύμφωνα’ (according to) have completely different meaning. The first word has positive polarity, whereas the second is neutral. Since the stem of both words is ‘συμφων’, there is no information to reveal the original word before stemming. As unigrams are used for predictions after stemming, the above case may be handled incorrectly. The described phenomenon can be frequently seen in Greek, even with words that are spelled exactly the same, but because they are stressed differently, their meaning changes. Note that these words are not homonyms as the word “like” for example, which serves both as verb and as proposition. In fact, purely homonyms with different polarity are extremely rare if non existent in the Greek language.

In order to handle the particular cases properly, a database is created with each record storing the following information: (i) stem, (ii) part-of-speech tag, (iii) polarity. The presence of unigrams (if any), on which the classifier based each prediction, are stored. If one of them, along with its part-of-speech tag, exists in the database, it is replaced with another one that has the same polarity. Specifically, if the unigram that exists in the database is positive, it will be replaced with the emoticon “:)”, whereas if it is negative, it will be replaced with the emoticon ":(". Finally, if the unigram is neutral, an article will replace it. At the moment, seven such cases are identified and stored in the database. However, this database can be continually improved by a user feedback mechanism.

The described particularities show that depending on language, different preprocessing steps may improve the performance of the classifier and thus it is not trivial to suggest a method that proves to be best for every language.

5 Experiments

There are two versions of the proposed method that are developed for the experiments. The first one uses SVM as the classification algorithm and we will refer to it as #Sentiment_v1. The second version is called #Sentiment_v2 and uses Logistic Regression. SVM uses linear kernel and the value of parameter C is 1.0. The implementations of both algorithms are provided by the Weka data mining software. The section of experiments is divided in two parts. The first part presents the results of the evaluation on the Greek dataset collected by us, whereas the second includes results of experiments on English data provided by SemEval 2013. In these subsections we also compare the proposed method to three pre-existing methods developed for English [5,9,17], which we followed and implemented according to the descriptions in the corresponding papers. We will refer to these methods as Go_method, Mohammad_method and Barbosa_method according to the first author. Due to space restrictions we do not describe these methods, but of course we provide the corresponding references for details.

The evaluation metrics we report in the experiments are average precision, recall and F-score, i.e. the sum of the corresponding metrics for each class divided by the number of classes. We also use McNemar's test [16] to check the statistical significance of the difference in performance between systems in each experiment.

5.1 Greek Data

The experiments of this section concern the evaluation on Greek data gathered by us, i.e. the GR–test. We present a comparison between the two versions of our system and the three pre-existing methods described above. For Barbosa_method we implement only two lexicon features, number of positive and number of negative words, as all other lexicon features depend on the structure of the MPQA lexicon [25], which is separated into strong subjective and weak subjective terms (this distinction does not exist in current Greek lexicons). Figure 1 displays the evaluation results of the five systems, #Sentiment_v1, #Sentiment_v2, Go_method, Mohammad_method and Barbosa_method on Greek data. As far as our method is concerned, the difference in F-score between #Sentiment_v1 and #Sentiment_v2 on GR–test is statistically significant, which means that SVM outperforms Logistic Regression. However, our performance is statistically indistinguishable from Mohammad_method. The other two methods by Go and Barbosa achieve much lower average F-scores.

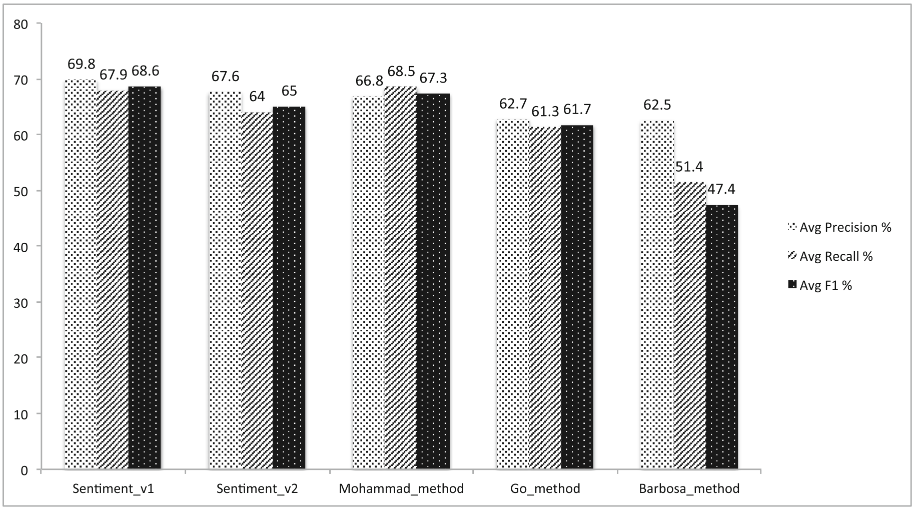

Fig. 1. Results on Greek data (GR–train for training and GR–test for testing)

The main conclusion of this experiment is that just the use of unigrams as features is not enough to achieve high accuracy in a classification problem with three classes. The Go_method was originally tested on a two-class classification of English tweets and generated good results, but the extension of the method to three classes and on another language seems not so simple and would need further preprocessing steps/features to work. This is demonstrated by #Sentiment_v1, #Sentiment_v2 and Mohammad_method, which also support the use of unigrams, but extend it with lexicon features, more preprocessing, such as stemming or feature selection, and achieve to reach higher average F-scores. Mohammad_method addresses inflection by keeping all ngram features, which however means a much larger model and more training time.

5.2 English Data

This section is dedicated to experiments on English data provided by the organizers of SemEval 2013. Again we present a comparison between the system proposed (only #Sentiment_v1, which is the best version according to the previous experiments) and the other three methods.

The evaluation results on the SemEval dataset (Sem–train and Sem–test) are displayed in Fig. 2. #Sentiment_v1 and Mohammad_method are again statistically indistinguishable and give the highest F-scores. Again methods that include both ngram and lexicon features, along with preprocessing and feature selection techniques, perform better on English data.

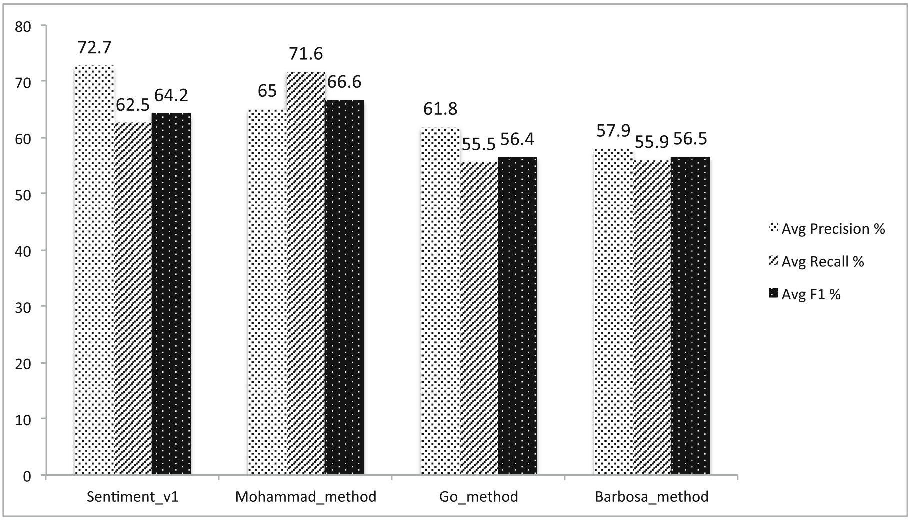

Fig. 2. Results on SemEval data (Sem–train for training and Sem–test for testing)

5.3 Time Consumption

In this section we present time performance results of the methods using the number of predicted tweets per second and training time. All methods ran on a single machine with an Intel Core i5 processor at 2.6 GHz and 16 GB of RAM. Details about time consumption of each method are given by Table 2.

Table 2. Time consumption

Method	Predicted tweets/sec	Training time (minutes)
#Sentiment_v1	16 tweets/sec	8.45 min
Mohammad_method	9 tweets/sec	14.91 min
Go_method	807 tweets/sec	5.9 min
Barbosa_method	8 tweets/sec	15 min

Although #Sentiment_v1 and Mohammad_method are indistinguishable in terms of F-score, #Sentiment_v1 needs 43 % less prediction and training time. This difference in time performance is reasonable, since Mohammad_method generates more features, such as part-of-speech and Twitter syntax features (RTs, hashtags, e.g.), which based on the experimental results they do not contribute that much to accuracy, but they increase processing time. Go_method is by far the fastest method. This is because it only involves unigram features, which are quickly generated. Nevertheless, they fail to predict test data effectively as experimental results demonstrated.

5.4 Sensitivity Analysis

We also performed ablation tests in order to check how the omission of different steps of our method affects performance. Table 3 shows the effect of negation identification, feature selection and stemming on Greek and English data. The remarkable change in F-score in Greek after the omission of stemming is expected due to the inflective nature of the language. Notably, negation identification does not seem to matter a lot. This is probably due to the fact that many tweets are neutral and their polarity is not reversed, but also that the technique suffers from low recall. It tends to be quite precise and correctly reverse polarity when a negation pattern is captured and the negated token is one of the classification features. However, in many cases the negated token does not belong to the features and even though the pattern is again captured, no reversal takes place.

Table 3. Results of sensitivity analysis

Modification	Avg F-score on Greek	Avg F-score on English
No modification	68.6 %	64.2 %
Without negation identification	68.7 %	64.1 %
Without feature selection	66.7 %	62.2 %
Without stemming	63.1 %	62.2 %

6 Conclusion and Future Work

We present a method for sentiment analysis in Twitter focused on the Greek language. We perform the first multilingual comparative analysis and report comparison results to three leading existing methods, from experiments on two different datasets (Greek and English). Our method clearly outperforms two of the three methods we compared against in sentiment extraction, while being statistically indistinguishable from the third. However, the proposed method needs 43 % less time for predictions and training. These experiments reveal that the generalization of a method to different languages or from a two to a three class classification problem is not trivial. Moreover, they give evidence about the effect of different preprocessing steps and features, such as stemming, in performance for Greek and English. An interesting idea to pursue in the future is the assignment of sentiment to the correct entity in the tweet.

References

1. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. **26**(3), 12:1–12:34 (2008)

2. Abdul-Mageed, M., Diab, M.T., Korayem, M.: Subjectivity and sentiment analysis of modern standard arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers . HLT 2011, vol. 2, pp. 587–591. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
3. Annett, M., Kondrak, G.: A comparison of sentiment analysis techniques: polarizing movie blogs. In: Bergler, S. (ed.) Canadian AI. LNCS (LNAI), vol. 5032, pp. 25–35. Springer, Heidelberg (2008)
4. Atteveldt, W.V., Ruigrok, N., Schlobach, S.: Good news or bad news? conducting sentiment analysis on dutch text to distinguish between positive and negative relations. J. Inf. Technol. Polit. **5**(1), 73–94 (2008)
5. 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. Association for Computational Linguistics (2010)
6. Boiy, E., Moens, M.F.: A machine learning approach to sentiment analysis in multilingual web texts. Inf. Retrieval **12**(5), 526–558 (2009)
7. Fleiss, J., et al.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378–382 (1971)
8. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. HLT 2011, vol. 2, pp. 42–47 (2011)
9. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Processing **150**(12), 1–6 (2009)
10. Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M.: Comparing and combining sentiment analysis methods. In: Proceedings of the First ACM Conference on Online Social Networks. pp. 27–38. COSN '13, ACM, New York, NY, USA (2013)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
12. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd International Conference on World Wide Web. WWW 2013 (2013)
13. Koleli, E.: A new Greek part-of-speech tagger, based on a maximum entropy classifier. Master's thesis, Athens University of Economics and Business (2011)
14. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: 1995 Proceedings of Seventh International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE (1995)
15. Lovins, J.B.: Development of a stemming algorithm. Mech. Translation Comput. Linguist. **11**, 22–31 (1968)
16. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
17. Mohammad, S., Kiritchenko, S., Zhu, X.: Nrc-canada: building the state-of-the-art in sentiment analysis of tweets. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2, pp. 321–327 (2013)

18. Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T.: Semeval-2013 task 2: sentiment analysis in twitter. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2, pp. 312–320 (2013)
19. Ntais, G.: Development of a Stemmer for the greek Language. Master's thesis, Stockholm's University (2006)
20. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA) (2010)
21. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
22. Petasis, G., Spiliotopoulos, D., Tsirakis, N., Tsantilas, P.: Sentiment analysis for reputation management: mining the Greek web. In: Likas, A., Blekas, K., Kalles, D. (eds.) SETN 2014. LNCS, vol. 8445, pp. 327–340. Springer, Heidelberg (2014)
23. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. EMNLP 2011 (2011)
24. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
25. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT 2005, pp. 347–354 (2005)
26. Zhao, J., Dong, L., Wu, J., Xu, K.: Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 1528–1531. KDD 2012 (2012)