A Hybrid Method for Sentiment Analysis of Election Related Tweets

Dimosthenis Beleveslis
The Data Mining and Analytics Research Group School of Science & Technology, International Hellenic University
Thessaloniki, Greece
d.beleveslis@ihu.edu.gr

Christos Tjortjis
The Data Mining and Analytics Research Group School of Science & Technology, International Hellenic University
Thessaloniki, Greece
c.tjortjis@ihu.edu.gr

Dimitris Psaradelis
InfiLab GmbH
London, United Kingdom
dimitrisps@infilab.io

Dimitris Nikoglou
InfiLab GmbH
Zurich, Switzerland
dimitrisnk@infilab.io

***Abstract*—**Political sentiment analysis using social media content has recently attracted significant interest. This paper focuses on analyzing tweets in Greek regarding the recent European Elections. A hybrid method that combines Greek lexicons and classification methods is presented. A probabilistic classification model, that predicts the sentiment of tweets, is combined with hashtag-based filtering. Based on the predictions, an analysis is implemented, that shows how the public sentiment was affected by specific events during the pre-election period.

***Keywords*—** *Sentiment Analysis, Natural language processing, Supervised learning, Classification, Social Media and e-technologies*

I. Introduction

Today, it's easy to find twitter content with people's views on various topics as they are strongly inclined to express online their opinion in social media. Sentiment analysis systems are applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors [1]. A domain that is commonly discussed in social media platforms is politics. As more and more users express their political views, twitter has become a valuable source of people's opinions and sentiments. It is very interesting to analyze such opinions as it can give a clue of what people think about the elections or other political events. Political parties may be interested to know if people support their campaigns or not [5], [12].

Greek language analysis is of great interest in the field of natural language processing, as it has not been yet analyzed in depth [2]. It can offer potential solutions to multiple issues that have arisen during the years of the rapid digital growth. While it is easy to find ready-made datasets and models for English, this is hard for Greek. One of the novelties of this paper is that we performed sentiment analysis in Greek text, as this has not been researched adequately, yet.

A typical approach to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In these lexicons, words are tagged with their polarity: out of context, does the word seem to evoke something positive or something negative? Each tweet is analyzed in a linguistic way in order to detect its sentiment. This is an unsupervised method of detecting the sentiment of a tweet. Another more sophisticated method is supervised learning. In this method a set of already rated tweets is necessary in order to build a model. The target is to build a model that can predict the sentiment of future unseen tweets.

In our paper, we focus on Twitter, the most popular microblogging platform, for the task of sentiment analysis. We use a dataset of rated tweets as a trainset of a supervised learning model. Furthermore, new tweets are gathered through the official twitter API. The model is based on features that are extracted through the text of the tweets and two Greek lexicons. Specifically, it is an attempt to detect the probability that a tweet belongs to one of three sentiment categories (positive, negative or neutral).

The model was used in order to analyze tweets related to the recent European Elections held in Greece on May 26th, 2019. In particular, we used the predictions that the model produced regarding tweets that we gathered during the pre-election period. The target was to detect the sentiment of Greek twitter audience during that period and how specific political events affected it. Taking into consideration that the Greek users seem to express more their negative opinions regarding politics, we focus on splitting the tweets in negative and non-negative ones.

The remainder of the paper presents related work in section II and discusses data collection and preprocessing in section III. Our method is detailed in section IV. Results are discussed in section V and the paper concludes with directions for future work in section VI.

II. Related Work

Over the past decade, sentiment analysis of text written in English has been analyzed in depth. Many researches were held and led to tools that can handle the issue satisfyingly [3]. In contrast, there have been limited research for the Greek language. However, during the last years there have been some interesting studies on sentiment analysis of Greek text. The approaches adopted could be grouped in two main categories: machine learning and semantic orientation approaches. The machine learning approach is a supervised task, as it involves the training of a classifier using a collection of representative data. On the other hand, the semantic orientation approach involves the determination of the document's overall sentiment from the semantic orientation of words that it contains, without prior training, thus an unsupervised method. Some of those researches were taken into consideration in our approach and are mentioned below.

A very broad overview of the existing work was presented by Tsakalidis et al. [2]. In their survey, the authors try to tackle the problems arising when analyzing text in such an under-resourced language. They presented and made publicly available a rich set of such resources, ranging from a manually annotated lexicon, to semi-supervised word embedding vectors and annotated datasets for different tasks. Some of these datasets were used in our research. They performed several experiments, using different algorithms on three different sentiment-related tasks and evaluated the resources.

Research was also conducted by Kalamatianos et al. on Sentiment Analysis of Greek Tweets and Hashtags using a Sentiment Lexicon [4]. This research focuses on Greek Tweets, investigating methods for extracting sentiment of individual tweets as well population sentiment for different subjects (hashtags). The proposed methods are based on a sentiment lexicon. The variation of sentiment intensity over time for selected hashtags was examined and associated with real-world events.

Promising research on predicting the winner of the US elections via sentiment analysis was attempted by Oikonomou and Tjortjis [5]. The analysis was in English, however, it is of great interest how sentiment analysis was used in this direction and also the positive results. In particular, we managed to predict that the winner of the election would be Trump despite the overwhelming sentiment in favour of Clinton to win.

Furthermore, Antonakaki et al. [6] published a relative survey regarding the elections for the Greek referendum in 2015 and the subsequent legislative elections. Novel dictionaries were compiled for sentiment and entity detection for the Greek was tailored to these events. They subsequently performed volume analysis, sentiment analysis, sarcasm correction and topic modeling. Results showed that there was a strong anti-austerity sentiment accompanied with a critical view on European and Greek political actions.

III. Data gathering and preprocessing

A. *Data gathering*

An important part of the survey was to gather the appropriate data. These are Greek lexicons and tweets data sets, on which the modeling phase and the final analysis was implemented. The data used are presented below.

1) Greek lexicon

A manually annotated Greek lexicon that was produced by Tsakalidis et al. [2]. It contains 32.884 words, manually annotated regarding their positivity, negativity and subjectivity. Words are also rated regarding 6 emotions (anger, disgust, fear, happy, sad, surprise). The scores range from 0 to 1. The preprocessing steps that were taken in order to create the Greek lexicon are presented in §III.B.1.

2) NGrams lexicon

Automatically generated keyword-based lexicon that was produced by Tsakalidis et al. [2]. It consists of 190.667 ngrams that are rated regarding their polarity. Specifically, there are 52.577 unigrams and 138.090 bigrams. The polarity score ranges from -15 to +15. The higher the score is the more positive the ngram is.

3) Rated tweets

It is hard task to find or create datasets with Greek rated tweets. A bottleneck of this research was the creation of such a dataset. For this reason, we used a dataset consisting of tweets related to the January 2015 Greek General Elections.

Finally, a dataset of 1.640 rated tweets was created that was used in the modeling phase. This dataset consists of 79 positive, 582 negative and 979 neutral tweets, as shown in Table I. It is obvious that the dataset is unbalanced as the positive tweets are much less than the negative and neutral. This characteristic is very intense in Greece in the domain of politics, as users tend to express more negative opinions regarding politics.

TABLE I. NUMBER OF RATED TWEETS PER CLASS

	Positive	**Negative**	**Neutral**	**Total**
Tweets	79	582	979	1.640

4) Unrated tweets

A dataset of new tweets and retweets was created by gathering them from twitter via the official twitter API that is available. The tweets were in Greek and they were searched based on specific hashtags related to the Greek 2019 elections. We also included hashtags related to the two most important political parties in Greece based on the previous elections: “SYRIZA” and “ New Democracy”. So, the hashtags were split in two main categories. The first corresponded to hashtags of general use that do not refer to a specific political party. The second category corresponded to hashtags that are connected to specific political parties or politicians. We chose to include only hashtags that were neutral and not those that were polarized by default. The most prominent hashtags that prevailed throughout the period that preceded the Greek elections are presented in Table II. This period was from 1st May 2019 until 21st May 2019. The elections took place on May 26th.

TABLE II. NUMBER OF RATED TWEETS PER HASHTAG

Hashtag	**Tweets**
#εκλογες (elections)	2.861
#Βουλή (parliament)	11.141
#Τσιπρας (Tsipras)	12.258
#Μητσοτακης (Mitsotakis)	7.909
#ΣΥΡΙΖΑ (SYRIZA)	8.613
#ΝΔ (New Democracy)	3.923
Total	46.705

B. *Preprocessing*

Preprocessing refers to the transformations applied to data before feeding them to the classification algorithm. Transforming text into something an algorithm can use is a complicated process. In order to achieve better results from the applied model, both the tweets and the lexicons have to be formatted properly. The preprocessing steps that were followed for the lexicons and tweets are described below.

1) Lexicons

The two Greek lexicons (see III. A.1-2) had to be formatted properly in order to be used for modeling. The main target was to transform the data in such a way that lexicon terms could be matched with those of the tweets. Below are the steps that we followed:

- Removal of accents – Greek words contain accents that can give different meaning to a word e.g. νόμος, νομός. It was decided to remove accents as many twitter users avoid using them in their tweets. In that way it was easier to match the terms of the tweets to the lexicon words.

- Capitalize words – All words of the lexicons were converted to uppercase. The same was done for the words of the tweets.
- Stemming – Stemming is a process where words are reduced to a root by removing inflection through dropping unnecessary characters, usually a suffix. Stemming was applied through python module “greek-stemmer” [7] in all terms of the Greek lexicon.

2) *Tweets*

Tweets contain hashtags, mentions, URLs and punctuation that had to be removed. There are also other important preprocessing steps in order to transform a tweet in a way that can be used to extract features out of it. It was also important to transform them in that way so that they could be combined with the lexicons. Below are the steps that were followed for this scope:

- Extract only text without hashtags, URLs and emojis - A tweet may contain many parts except its text. All these parts had to be removed in order to keep only the plain text. For this scope the python package “tweet-preprocessor” [8] was used.
- Remove accents - As mention in paragraph III.B.1 Greek words contain accents that were removed.
- Capitalize words – Following the same procedure as in the lexicons, all terms of the tweets were converted to uppercase.
- Remove punctuation - Punctuations do not help much in analyzing the sentiment. So, they were removed.
- Tokenize - Tokenization is a step which splits longer strings of text into smaller pieces, or tokens. Creating tokens out of tweets was necessary in order to match the words of the tweets to the lexicon words.
- Bigrams creation – Except from single tokens (unigrams), all possible bigrams were extracted for each tweet. A bigram is a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words. A bigram is an n-gram for n=2 [11].
- Stemming - As mention in paragraph III.B.1 each word was reduced to a root by removing its suffix.

C. *Feature generation*

Feature generation is a significant step before applying any machine learning algorithm. The quality of the results of those algorithms depends on the quality of the available features. Algorithms require features with some specific characteristics in order to work properly. Features were extracted from both the text of the tweets and the two lexicons.

1) *Tweets based features*

The features extracted from tweets are presented below.

- Number of exclamation marks: An exclamation mark may indicate positive or negative sentiment. Their number was calculated for each tweet.
- Number of question marks: With the same logic, the number of question marks was calculated.
- Number of hashtags and mentions: The number of hashtags and mentions that each tweet contains was also calculated.
- Number of full capitalized words: A fully capitalized word may indicate positive or negative sentiment. The number of them was calculated for each tweet.
- Number of words that each tweet consists of.
- A binary attribute that indicates whether the tweet contains a URL or not.
- Number of nouns, verbs and adjectives: By using the python package spacy [9], the number of nouns, verbs and adjectives was calculated.

2) *Lexicon based features*

The features that were created by combining the two lexicons with the preprocessed tweets are presented below.

- The number of positive and negative unigrams that appear in the Ngrams lexicon; the same for bigrams.
- The total summation of the scores of positive and negative unigrams that appear in the Ngrams lexicon; the same for bigrams.
- The difference between the overall scores of the positive and the negative ngrams.
- The sum of the scores regarding subjectivity, positivity and negativity from the Greek lexicon.
- Sum of the scores regarding six emotions (anger, disgust, fear, happiness, sadness, surprise) from the Greek lexicon.

IV. METHODOLOGY

The aim of this research was to build a model that predicts the sentiment of a Greek tweet. The problem is transformed to a classification problem that consists of three classes: negative, neutral and positive. So, the final model can classify each tweet based on its features, as described above. For this purpose, a probabilistic classification method is applied with the aim to build a model that can predict the probability that a tweet belongs to one of the three classes. Based on this outcome, each tweet is finally classified to one of the three classes. Moreover, besides the above supervised learning methods, a hashtag-based filtering is applied to each tweet. The objective of this process is the development of a model, which can be used for the prediction of future tweets.

A. *Classification*

In the modeling phase, various classification algorithms were applied. The most important are Random Forest, Decision Tree and XGBoost. As described in paragraph §III.A.3, the training set consists of 1.640 already classified tweets. This dataset was split in train set and test set with 10-fold cross validation. In addition, we performed parameter tuning for every algorithm in order to detect the best combination of parameters. Finally, the best scores were produced using Random Forest. Evaluation metrics for the algorithms we used are presented in Table III. It is obvious that the model is capable of predicting neutral and negative tweets well. However, it has a difficulty in predicting positive tweets. As we discussed in paragraph §III.A.3, the positive tweets are much less than the negative and neutral ones, in the domain that we are examining. The confusion matrix for the model using Random Forest is shown in Fig 1.

TABLE III. CLASSIFIERS USED AND EVALUATION METRICS

Algorithm	**Accuracy**	**Precision**			**F-Score**		
		Neg	***Neut***	***Pos***	***Neg***	***Neut***	***Pos***
Random Forest	0.80	0.74	0.83	1	0.73	0.83	0.34
Decision Tree	0.70	0.62	0.81	0.17	0.63	0.78	0.24
XGBoost	0.79	0.73	0.83	0.43	0.71	0.85	0.35

B. Probabilistic classification

A probabilistic classifier is a classifier that can predict, given an observation of an input, a probability distribution over a set of classes, rather than only outputting the most likely class that the observation should belong to.

A simple multiclass classification model as described in IV.A classifies each tweet to one of the three classes (negative, neutral, positive). However, a tweet usually contains more than one sentiment. Many tweets consist of words that refer to both positive and negative sentiments. So, it would be useful to find out to which extent each tweet belongs to each of the above classes. It would be interesting to know that a tweet is 90% negative and another tweet is 55% negative. We could say that the first tweet is definitely negative, but we cannot be sure about the second. Such a model is called a probabilistic classification model. Each of the algorithms mentioned in IV.A are transformed to a probabilistic model by using the package ‘scikit-learn’ [10].

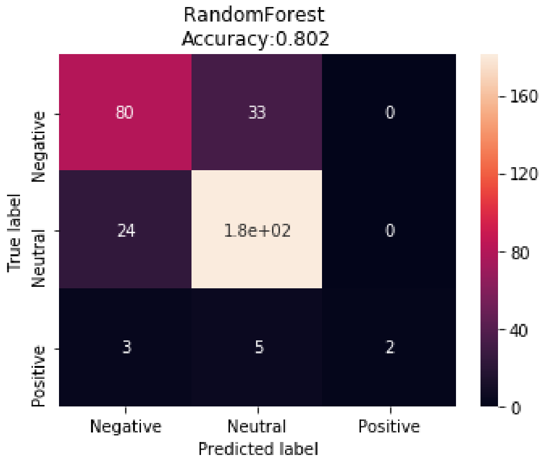

Fig. 1. Confusion matrix for random forest

The outcome of this probabilistic classifier was used in order to classify each tweet. Specifically, a tweet was considered to be negative if the probability to be negative was higher than 0.65. In any other case the tweet was considered to be non-negative. The distribution of the probabilities that each tweet is negative is presented in Fig. 2.

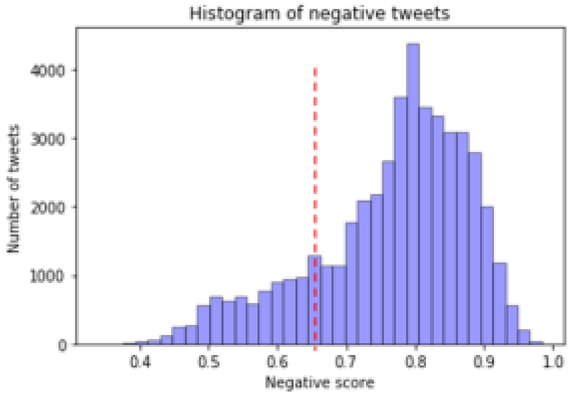

Fig. 2. Distribution of the probability that each tweet is negative

C. Hashtag-based filtering

In addition to the above procedure, a hashtag-based filtering was applied. After analyzing the tweets gathered, specific popular negative hashtags were found. These hashtags imply a negative sentiment. So, tweets that include one of those hashtags were classified as negative. The popular negative hashtags along with the number of tweets that they occur in, are presented in Table IV.

TABLE IV. NUMBER OF TWEETS CONTAINING EACH HASHTAG

Hashtags	Tweets
syriza_xeftiles	2.899
ndxeftiles	2.085
skai_xeftiles	753

V. Results

A. Predictions

The gathered tweets were classified by using the probabilistic classification model and the hashtag-based filtering described above. Taking into consideration that in the specific domain of Greek politics, the number of positive tweets is very small, the tweets were categorized as negative and non-negative. Finally, as shown in Fig. 3, in a set of 46.705 tweets and retweets that were gathered during the pre-election period in May 2019, there were 33.681 negative and 13.024 non-negative tweets. The tweets were searched based on specific hashtags mentioned in III.A.4. These belong to two categories: tweets obtained by searching election related neutral hashtags and tweets with hashtags related to the two most important political parties and their leaders.

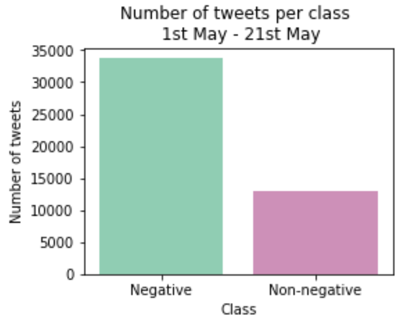

Fig. 3. Number of tweets per class

B. Analysis

Based on the predictions that are presented in V.A, an analysis was implemented. The target was to observe the reactions of the Greek twitter audience during the pre-election period in May 2019. Specifically, we tried to identify the changes in user responses during the pre-election days and link them to specific political and non-political events. The period examined is from 1st to 21st of May, while the elections were held on the 26th of May. We focus on the two most important, based on the previous elections, political parties in Greece and their leaders. These parties are SYRIZA and New Democracy, while their leaders are Tsipras and Mitsotakis respectively. This analysis was published as a blog article- [12]~~[12]~~.

1) Number of tweets per day

Although the tweets' volume is not a sufficient indicator of political inclinations of users, it can give insights regarding specific events. In Fig. 4 we plot the volume of tweets per day during the pre-election period of 21 days. We include tweets of all hashtags that are mentioned in III.A.4. The spikes in this plot are indicative of major events during that period. Analysis of the text from these tweets revealed the events that are presumably the cause of changes in the rate at which users were commenting on the elections.

Firstly, the increase in tweets appears after the May Day that followed Easter celebration. This first increase is due to the intense debate of the days about the tragedy in Mati that took place in July 2018. A sharp increase in comments was observed during 6-8 May, where the main topics of discussion were the presence of Tsipras in the yacht of a businessman and his announcement of positive economic measures. The vote of confidence for the government, as well as the speeches of political leaders in Greece where strongly discussed. Specific non-political events seem to have distracted the audience attention from the elections.

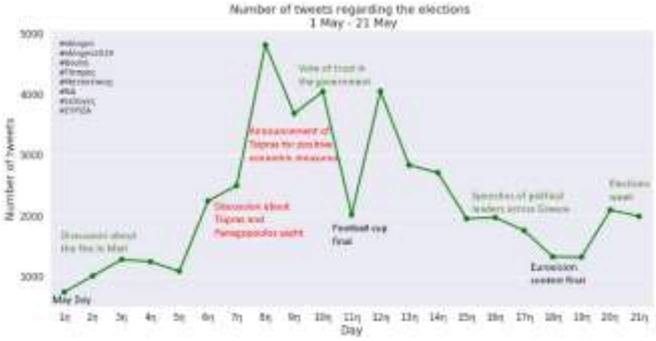

Fig. 4. Number of tweets per day regarding the elections

2) Number of tweets for SYRIZA vs. ND

In the same sense, it is interesting to compare the number of comments referring to the two largest political parties, according to the results of the previous 2015 elections. Based on Fig. 5, it seems that SYRIZA, that is the government, receives most of the comments almost every day. However, in the middle of the period, New Democracy received many more comments and this seems to be due to a statement by their leader about seven-day working week. In Fig. 6, we see that in the case of NEW DEMOCRADY, tweets related to their leader were more than those related to his party as a whole. The exact opposite happened in the case of SYRIZA.

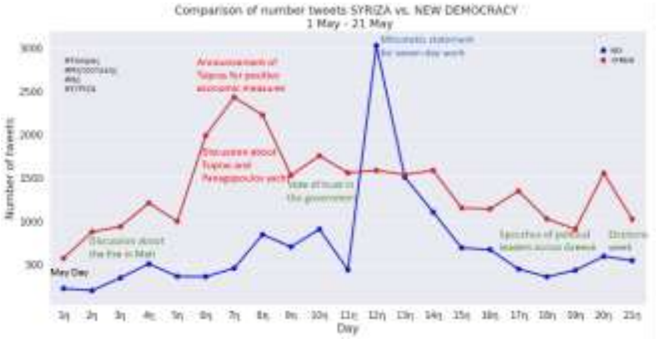

Fig. 5. Number of tweets per day related to ND and SYRIZA

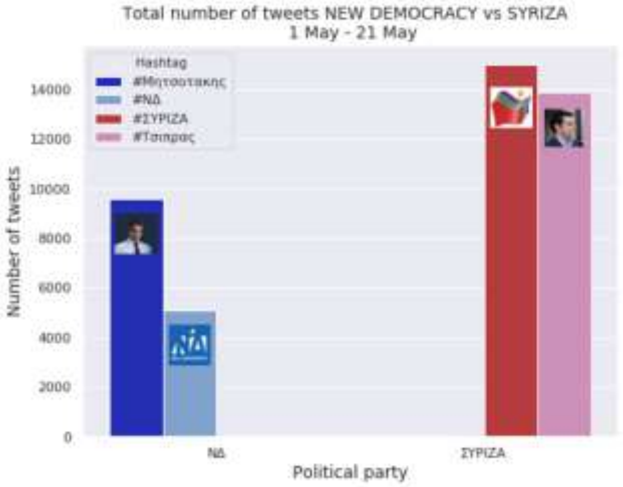

Fig. 6. Total number of tweets related to ND and SYRIZA

3) Comparison of negative tweets for SYRIZA vs. ND

Apart from the number of tweets, it is important to observe the percentage of tweets that were negative towards the two parties. Based on Fig. 7, SYRIZA receives more negative tweets almost every day with minor exceptions. A positive observation for New Democracy is the fact that there are days they receive the lowest rates of negative feedback, but this is associated with the much lower number of tweets that appear to be relevant to them. Specific events seem to have influenced the feelings of the users. Perhaps these have resulted in Mitsotakis gathering more negative comments than the party as a whole, as we can see in Fig. 8. On the other hand, Tsipras, the Prime Minister seems to receive fewer negative comments than SYRIZA as a whole. In any case, the strong tendency of Twitter users to comment with a negative mood is apparent.

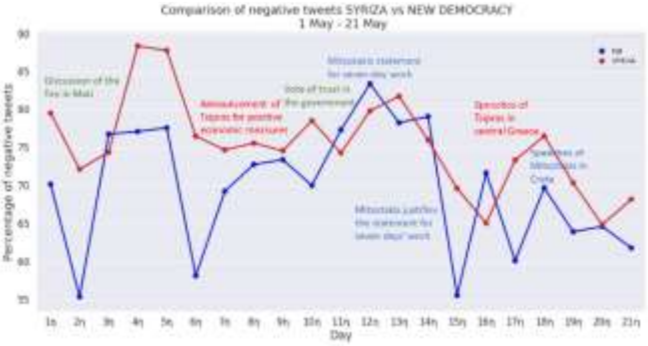

Fig. 7. Percentage of negative tweets per day related to New Democracy and SYRIZA

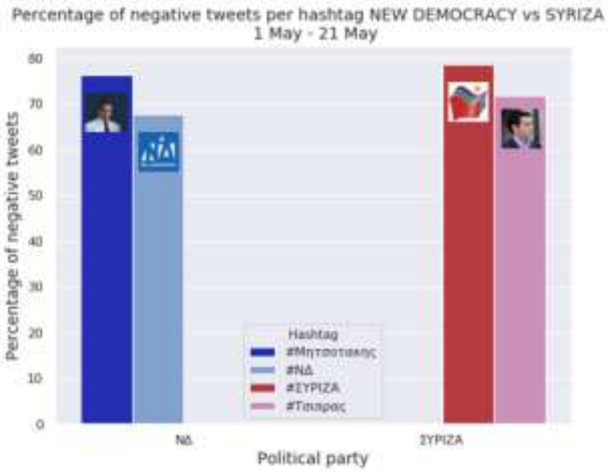

Fig. 8. Percentage of negative tweets related to New Democracy and SYRIZA in total

4) *Percentage of negative tweets per hashtag*

The intense phenomenon of negative election-related comments is clearly presented in Fig. 9. For both parties and political leaders, more than 65% of the tweets have negative sentiment. High percentages of negative tweets also appear in hashtags that are not related to specific parties, but generally to the elections. This is a very important insight and it might be an indicative of the high percentage of abstention in the elections in Greece.

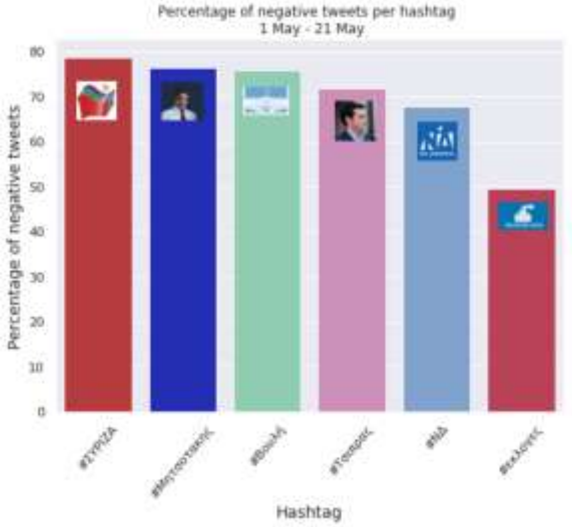

Fig. 9. Percentage of negative tweets related to popular hashtags

VI. Conclusions and Future Work

The purpose of this work was to create a sentiment analysis model that can analyze tweets related to politics written in modern Greek. This was accomplished based on classification methods and by using older rated datasets of tweets and Greek lexicons. Data was preprocessed accordingly and features were created in order to find the best classification algorithm for this approach. Furthermore, a probabilistic classification and a hashtag-based filtering were used in order to classify each tweet as negative or non-negative. A new dataset was created, consisting of tweets related to the Greek elections that were held in 26th of May 2019. These tweets were searched and gathered based on popular hashtags. The sentiment of the tweets was analyzed based on the above procedure. Finally, based on the predictions an analysis was implemented in order to observe the changes in user responses during the pre-election days and link them to specific political and non-political events.

In the future we plan to extend the research to even more sophisticated methods. A very interesting approach is to use Greek word embeddings and apply neural networks. Furthermore, it is definitely worthy to create a larger and more representative dataset of rated tweets that will help in the training of the model. Additional feature engineering can also improve the accuracy of the model. Finally, the combination of the current approach with a sarcasm detection model is of great interest.

References

[1] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol. 2012;5(1):1–67.

[2] Tsakalidis, A., Papadopoulos, S., Voskaki, R., Ioannidou, K., Boididou, C., Cristea, A., Liakata, M., & Kompatsiaris, I., (2018). “Building and evaluating resources for sentiment analysis in the Greek language”, *Language Resources and Evaluation*, Vol. 52, No. 4, pp 1021–1044.

[3] Z. Madhoushi, A. R. Hamdan and S. Zainudin, "Sentiment analysis techniques in recent works," 2015 Science and Information Conf. (SAI), London, 2015, pp. 288-291.

[4] Kalamatianos, G., & Mallis, D., Symeonidis, S,, Arampatzis, A.,. (2015). “Sentiment Analysis of Greek Tweets and Hashtags using a Greek Sentiment Lexicon”, *Proc. 19th Panhellenic Conf. Informatics* (PCI '15), pp. 63-68.

[5] Oikonomou, L. & Tjortjis, C. (2018). “A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter”, *3rd IEEE SE Europe Design Automation, Computer Engineering, Computer Networks, and Social Media Conf.* (IEEE SEEDA-CECNSM18).

[6] Antonakaki, D., Spiliotopoulos, D., Samaras, C., Pratikakis, P., Ioannidis, S., Fragopoulou, P., (2017). Social media analysis during political turbulence, PLoS ONE 12(10): e0186836.

[7] greek-stemmer 0.1.1, pypi.org/project/greek-stemmer/

[8] tweet-preprocessor 0.5.0, pypi.org/project/tweet-preprocessor/

[9] Spacy, Industrial-Strength NLP in Python, spacy.io/

[10] scikit-learn, Machine Learning in Python, scikit-learn.org

[11] [Wikipedia contributors]. "Bigram." Wikipedia, The Free Encyclopedia.

[12] ‘Reactions of Greeks in twitter regarding the political events during the last days’, InfiLab GmbH blog infilab.io/blog-2/