Monitoring the Twitter sentiment during the Bulgarian elections

Jasmina Smailović, Janez Kranjc, Miha Grčar, Martin Žnidaršič, Igor Mozetič
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

***Abstract*—We present a generic approach to real-time monitoring of the Twitter sentiment and show its application to the Bulgarian parliamentary elections in May 2013. Our approach is based on building high quality sentiment classification models from manually annotated tweets. In particular, we have developed a user-friendly annotation platform, a feature selection procedure based on maximizing prediction accuracy, and a binary SVM classifier extended with a neutral zone. We have also considerably improved the language detection in tweets. The evaluation results show that before and after the Bulgarian elections, negative sentiment about political parties prevailed. Both, the volume and the difference between the negative and positive tweets for individual parties closely match the election results. The later result is somehow surprising, but consistent with the prevailing negative sentiment during the elections.**

I. Introduction

Recently, as a consequence of massive use of Internet and new technologies, an explosive growth of user-generated content is observed. Using on-line websites, such as blogs, forums and social networking platforms, people write about their observations, opinions and emotions. They comment on various subjects – individuals, companies, political parties, movements or popular events. Several of these on-line platforms have become very popular, for example, the social networking and micro-blogging platform Twitter (http://www.twitter.com), which attracted over 500 million active users and generates 500 million tweets daily [20].

Massive amounts of user-generated content represent a relevant source for gathering and analyzing people's viewpoints, opinions and emotions. Consequently, many researchers are interested in such data. Data from social network sites (e.g. Twitter, Facebook, etc.) is especially interesting for examination because of large volumes and near real-time posting.

There are reports on discoveries of a number of relations between the data from social networks (e.g., the volume, sentiment polarity) and various phenomena of interest (e.g., stock prices, products' success, election results). These studies are often controversial and the methodologies and software tools are in the early stages of their development and common acceptance. In this paper, we present a particular study on monitoring and analyzing the Twitter sentiment during the Bulgarian elections. We describe in detail a novel approach to real-time monitoring of Twitter, along with the conducted experiments and evaluation results.

A. Motivating use case

Our first attempt at real-time monitoring of the Twitter sentiment was during the Slovenian presidential elections in 2012. In collaboration with a TV station POP TV, we have developed a sentiment analysis platform, which collected and analyzed tweets about the three presidential candidates. The sentiment charts were shown in prime time during live TV debates on POP TV. The results were highly controversial in the sense that they were in conflict with the polls carried out by the major polling agencies. All polling agencies predicted the Slovenian president at the time (Mr. Danilo Türk) to win the first round of elections, while our system clearly showed the lead of Mr. Borut Pahor. Eventually, Mr. Borut Pahor won both rounds of elections and is the current Slovenian president.

During the TV debates, the volume of tweets about the three candidates, representing the viewers engagement, was nearly ten times higher than at a comparable time period during the rest of the campaign (excl. the two election days) and approximately two times higher than at times when the presidential debates were aired on the Slovenian national television. In addition, POP TV noted that the ratings of their shows were three to four times higher than those of the comparable competition and that there was no naturally occurring audience decline during the shows. This gives an indication that feedback through Twitter is not just a marginal public opinion indicator but has a potential for increasing public engagement and providing early signs of prevailing public sentiment.

In April 2013, one month before the elections in Bulgaria, we were approached by a Bulgarian partner to conduct a similar real-time monitoring of the Twitter sentiment. This paper presents a first study of this kind done on Bulgarian tweets. In the process, we made several improvements to the technology and methodology for real-time monitoring of the sentiment in tweets: an improved labeling mechanism, a novel definition of the SVM classifier's neutral zone, an improved language detection method, and an extensive evaluation.

B. Related work

Several studies have analyzed sentiment in Twitter posts and its relation to trends. In this section we first present some studies about the Twitter sentiment analysis in general, and then focus on studies that were targeting the political domain and elections. Many studies in this field deal with U.S. elections and they are introduced first, followed by studies concerned with elections in other countries.

Analysis of sentiment proved to be an interesting and useful tool in a very diverse array of domains and applications. Thelwall et al. [41] showed that there is a relationship between popular events in Twitter and increases in negative sentiment. In [23], the authors focused on target-dependent sentiment classification and applied it to English tweets containing popular topics. Asur et al. [2] employed tweet-rate about specific topics to build a model for predicting box-office profit of movies in advance of their release. An application of sentiment analysis additionally improved their results. Bollen et al. [4] measured mood in tweets in terms of six dimensions (calm, alert, sure, vital, kind and happy) and showed that changes in calmness can predict daily changes in the closing values of the DJIA index. In our previous work, the volume and sentiment of financial tweets were used to identify important events, and changes in positive sentiment probability were used as indicators of future stock price movements [37], [38]. We have also showed that Twitter sentiment has significant impact on the DJIA stock prices [33].

Elections are phenomena that usually stir a lot of attention and (emotional) response and the election results are one of the better documented reflections of public mood. There has been a lot of research on this topic, particularly on the question whether the analysis of social media can be used to predict the outcome of elections. A survey is given by Gayo-Avello [17]. Conclusions are different: from those claiming that data from social media is a reliable predictor (one of the first were the papers by O'Connor et al. [30] and Tumasjan et. al [44]), to those presenting the opposite findings.

O'Connor et al. [30] analyzed correlations between the public opinion in U.S. from polls, and Twitter. The authors used a simple method for estimating the sentiment in tweets by looking at a presence of positive and negative sentiment words from a subjectivity lexicon. On one hand, they found that sentiment in Twitter messages does not substantially correlate to the U.S. presidential election polls in 2008, but on the other, they showed a considerably high correlation with the index of Presidential Job Approval. Also, they found that message volume for "Obama" had high correlation to the polls, but the same was observed also for "McCain" and Obama's ratings. In [8] the authors used data from the 2010 U.S. Senate elections in Massachusetts and applied a prediction method, which uses share of tweets for each candidate, as in [44], and a method which calculates sentiment in tweets, as in [30]. The authors argued that studies which had shown a direct correlation between volume/sentiment of Twitter data and outcome of elections had many shortcomings and that their methods were no better than random classifiers. Similarly, Gayo-Avello et al. [15] used the somewhat modified approaches of [44] and [30], and examined the predictive power of Twitter data during the 2010 U.S. Congressional elections. They found no correlation between the analysis and the election results, contradicting previous reports. In the other paper [16], Gayo-Avello analyzed in detail the reasons for failing to predict the results of the 2008 U.S. elections and provided several lessons that can be learned from this research. The authors in [29] calculated predictions for two 2010 U.S. Congressional elections based on the share of tweets for each candidate, as in [44], and sentiment in tweets, similar to [30]. Their experiments showed that the data from social media did only slightly better than chance in predicting election results. Livne et al. [27] analyzed tweets posted by the candidates during the midterm U.S. elections in 2010. Using different linear regression models, where independent variables were graph properties, Twitter-derived variables and candidate's and party's properties, the authors reported 88% accuracy when using Twitter-derived variables over 81% without them. The authors in [10] used Twitter data and data from competitive 2010 U.S. congressional election outcomes and showed that the percentage of Republican-candidate name mentions correlates with the Republican vote margin. Finally, the authors in [22] analyzed Twitter data in relation to the 2012 U.S. presidential elections and showed that such analysis can reveal how popular the candidates are. Furthermore, the authors demonstrated that the results for individual U.S. states can be predicted by employing sentiment analysis and tweets' geographical information.

In what follows, we present studies focused on the analysis of elections in some other countries. Tumasjan et al. [44] showed that Twitter was heavily used as a platform for political discussion regarding the 2009 German federal elections. The authors demonstrated that the mere count of tweets mentioning a certain party reflects the election outcome, while the sentiment of Twitter messages closely corresponds to the offline political landscape. This work triggered many discussions. Some studies criticized the proposed approach and reported its shortcomings (e.g., [8], [15], [25], [26]), while the others supported it (e.g., [5], [45]). Bermingham and Smeaton [3] developed a system which provided a real-time interface into Twitter discussions about the 2011 Irish General Election. The authors showed that both volume-based measures and sentiment analysis have predictive power, with the volume being a stronger indicator than sentiment. However, it was also reported that the developed methods are not competitive with the standard polling approaches. Borondo et al. [5] analyzed Twitter data during the 2011 Spanish presidential elections, and found correlation between the user activity and the election results. They supported the approach by [44], as they showed that relations in votes and tweets between the two main political parties in Spain reasonably strongly correlate. Sang and Bos [34] analyzed Twitter data in relation to the 2011 Dutch Senate elections and employed the prediction method from [44]. Their results showed that the number of tweets that mention political parties is not a good predictor and that the performance can be improved by applying sentiment analysis. Skoric et al. [35] tested the predictive power of tweets in the 2011 Singapore general elections. They showed that there is moderately strong correlation between the share of tweets and the share of votes at the national level. At the level of constituency, this correlation is much weaker. The accuracy of the predictions in this research was significantly lower than the one reported by Tumasjan et al. [44]. Caldarelli et al. [6] analyzed tweets and their volume per political party in the context of the 2013 Italian national elections. Their experiments show that the tweet volume and its changes in time can be used as indicators of the final election outcomes at the national level and macro areas. Finally, Eom et al. [11] analyzed the volume of tweets during two elections in Italy, and one in Bulgaria (the same elections as analyzed in this paper). Their results are consistent with ours, and show that the tweet volume can indicate election results if the optimal period of averaging the volume is taken into account.

The overview of related work indicates that a lot of research

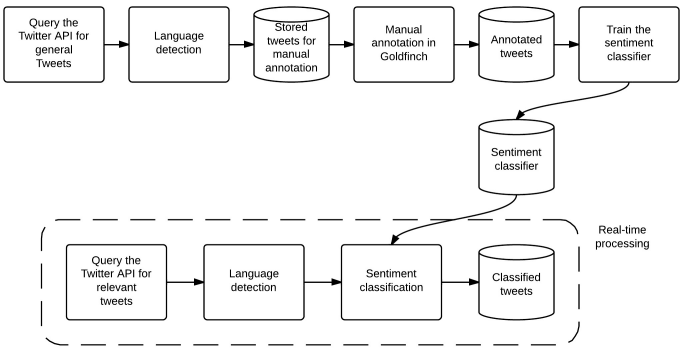

Fig. 1. A flowchart of the sentiment classifier training and the real-time tweet classification processes.

was dedicated to the analysis of correlation between the sentiment in Twitter posts and election outcomes in various scenarios, but the conclusions are varying, sometimes even contradictory.

C. Overview

Most of the former research on predicting the outcome of elections from Twitter has been based on the tweet volume or simple sentiment analysis techniques, i.e., counting positive and negative sentiment words in tweets. In this paper, we employ supervised machine learning and use the Support Vector Machine (SVM) algorithm on a large training set. The methodology that we propose is generally applicable to various real-time Twitter monitoring scenarios.

The generic process of monitoring the Twitter sentiment is presented in Figure 1: (i) training the sentiment classifier, and (ii) applying the classifier to real-time data. Training the classifier (the top row of Figure 1) is a time consuming process since manual annotations of a large number of tweets are required. The classification process (the bottom row of Figure 1) is executed for each incoming tweet, and has to satisfy the real-time constraints, i.e., the rate of processing needs to be higher than the rate of incoming tweets. For subsequent visualizations, the classified tweets are stored in a database where they can be efficiently processed.

The paper is structured as follows. Section II is the core of the paper. We address all the ingredients needed to perform the Twitter sentiment analysis: the annotation process, preprocessing of tweets and feature selection, the SVM classifier construction, introduction of the neutral zone, and extensive evaluation. In Section III we analyze the Bulgarian election results and compare them to the Twitter sentiment before and after the elections. In Section IV we show the problems with the Twitter language detection and construct a considerably improved language classifier for Bulgarian tweets. We conclude with ideas for further work in Section V.

II. Sentiment Analysis

The core idea of our approach to sentiment analysis is to automatically learn classification models from manually annotated data. This requires relatively costly engagement of qualified annotators, but, on the other hand, it results in high quality sentiment classification models. If the tweets are domain specific, and not general, a lower number of annotated tweets results in a higher quality sentiment classifier. This was observed during our monitoring of the Slovenian presidential elections, and in a work on sentiment leaning towards environmental topics [36]. For this study, we were able to annotate general tweets (about 30,000), because only a relatively low number of political tweets were available before the Bulgarian elections. We trained several classification models on general tweets, and applied the final one to political tweets in real-time during the elections.

Another issue is the treatment of tweets for which there is no clear sentiment polarity (positive or negative), is too difficult to determine, or the tweets are irrelevant for the subject. We ignore such tweets (labeled as neutral) during training, and produce a binary (positive vs. negative) classifier. However, during classification, we introduce a measure of reliability, and tweets which are labeled as positive or negative below the reliability threshold are classified as neutral. By varying the reliability threshold, we can narrow or widen the neutral zone, and can find an appropriate compromise between the sensitivity (true positive rate) and specificity (1−false positive rate) of predictions.

We evaluate the constructed sentiment classifiers in three ways: by 10-fold cross validation, on the gold standard data, and by comparison to the actual elections results. The first two evaluations give the estimated sentiment accuracy of between 75% and 80% which is approaching, but not reaching, the inter-annotator agreement. A comparison to the actual elections results in Bulgaria is in Section III.

A. Twitter corpora

This section presents the Twitter data collected for the purpose of our study. We used the PerceptionAnalytics platform (http://www.perceptionanalytics.net) which allows the user to monitor public opinion about arbitrary subjects. The platform collects all relevant results from various social media sources (such as Twitter, Facebook, etc.), performs the analyses, and displays the results via its web interface.

We have collected two sets of Twitter data:

- *General* Bulgarian tweets (29,433), in the period from April 16 to April 29, 2013. The selection criteria were the geolocations of larger Bulgarian cities. These tweets were used for manual sentiment annotation, training of the sentiment classifiers, selection of appropriate features, evaluation by cross validation, and eventual classifier application during the real-time monitoring of political sentiment.
- *Political* Bulgarian tweets (10,300), in the period from April 29 to May 15, 2013. These tweets are the result of real-time monitoring of Twitter sentiment before and after the Bulgarian elections (held on May 12, 2013). The selection criteria were the names of major Bulgarian political parties and their leaders. The tweets are used to compare the sentiment and volume of the individual political parties to the actual elections results. A small subset of this dataset was also hand labeled and used as a gold standard in evaluation.

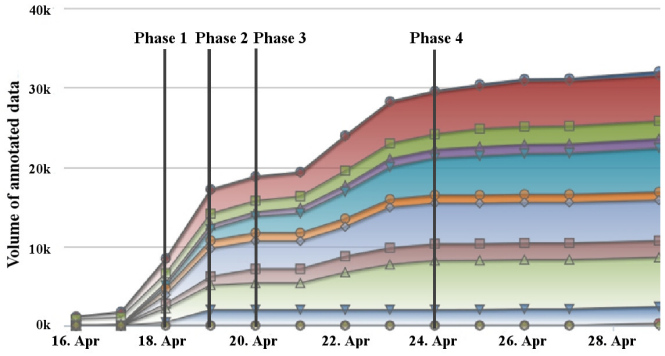

Fig. 2. The growth of annotations from twelve annotators through time (between April 16 and April 29, 2013). Phases after which new sentiment classifiers were constructed are marked with vertical lines.

B. Manual sentiment annotation

The general Bulgarian tweets were annotated using the Goldfinch platform, provided by Sowa Labs (http://www.sowalabs.com). Annotating the data is a process of assigning predefined labels to data instances. Twelve annotators were engaged to label the Bulgarian tweets as negative, neutral or positive. Tweets which were considered inappropriate or irrelevant, were excluded. The 29,433 general tweets were annotated in four phases (see Figure 2); 3,258 were excluded, and the rest has the following sentiment distribution: 6,716 negative, 11,015 neutral, and 8,444 positive (see Figure 8).

After each annotation phase, a new sentiment classifier was trained, evaluated by cross validation, and the preprocessing settings were selected which maximized the prediction accuracy. A training example (a tweet), with a sentiment label, is described by a vector of features which are words or n-grams that appear in the tweet. In the following subsections, we describe the algorithm used to train the sentiment classifiers, and different tweet-specific preprocessing settings.

C. Sentiment classification algorithm

There are three common approaches to sentiment classification [32]: (i) machine learning, (ii) lexicon-based methods and (iii) linguistic analyses. Instead of developing a Twitter-specific sentiment lexicon, we used a machine learning approach to learn a sentiment classifier from a set of class labeled examples. We adapted a linear SVM algorithm [46], standardly used in text mining. SVM has several advantages: it is fairly robust to overfitting, it can handle large feature spaces and it is memory efficient. Given a set of positive and negative training examples, the SVM training algorithm builds a model which separates the examples by a hyperplane. New unlabeled examples are projected into the same feature space and assigned a class label depending on the side of the hyperplane. The linear SVM decision function has the following form:

$$d(x) = x \times w + b = x_1 \times w_1 + x_2 \times w_2 + ... + x_n \times w_n + b \quad (1)$$

where x is the TF-IDF feature vector of a document to be classified, w is the SVM weight vector, and b is the hyperplane bias. TF-IDF stands for term frequency-inverse document frequency feature weighting scheme where weight reflects a relative importance of a word in a document collection.

The core components of Twitter data processing and classifier training are implemented by the LATINO library (Link Analysis and Text Mining Toolbox, http://latino.sourceforge.net). LATINO implements several SVM classifiers, the one used to train our tweet sentiment classifier is a wrapper around the SVM^{light} implementation [24].

D. Twitter data preprocessing

Preprocessing is a very important step in data analysis, particularly when textual data is used. We first performed the standard text preprocessing [13] to define the feature space for the training feature vector construction. These include text tokenization, removal of stopwords (e.g., and, or, a, an, the, ...), lemmatization and n-gram construction (n varies from 1 to 3). We did not use any part-of-speech (POS) tagger, since it was indicated by Go et al. [18] and Pang et al. [31] that POS tags are not useful when using SVMs for sentiment analysis. Moreover, we were not aware of a POS tagger, specific for tweets in Bulgarian, at the time of this study.

The resulting terms were used as features in the construction of the TF-IDF feature vectors representing the tweets. Additionally, tweets have some peculiarities which have to be taken into account. In preprocessing, we considered the following tweet-specific features [1], [18], [39]: *Usernames*, i.e., mentions of other users (e.g., *@john*), can be removed or replaced by a unique token *USERNAME*; *web links* can be removed or replaced by a unique token *URL*; *stock symbols* (e.g., *$AAPL*), can be removed; *letter repetitions* in a word can be replaced by only one occurrence of the letter (e.g., *looooove* replaced by *love*); *exclamation and question marks* can be replaced by tokens *EXCLAMATION* and *QUESTION*, respectively.

We experimented with these options to find the best combination of them. The preprocessing experiments were conducted on about 20,000 hand-labeled tweets, after the phase 3 of the annotation process (see Figure 2).

E. Evaluation by cross validation: the two class problem

For classifier training, only negative and positive tweets were used, and a binary SVM classifier was constructed. The best preprocessing setting was determined according to the averaged accuracy of the 10-fold cross validation. In our case, the setting uses 1- and 2-grams, terms with minimum frequency 1 in the corpus, replaces usernames with the token *USERNAME*, replaces web links with the token *URL*, removes stock symbols and removes repetitive letters. We calculated the 10-fold cross validation accuracy on negative and positive tweets after each phase. The results are in Figure 3. As expected, the highest accuracy (76.1%) is achieved in the last phase where the largest number of hand-labeled tweets is available (29,433). We observe the raise of accuracy with larger training sets, and diminishing improvements. As a consequence, we did not proceed with the manual annotation of additional tweets after phase 4.

F. Introducing the neutral zone

In this section we show how a binary SVM classifier can be extended to address the classification of tweets into three sentiment classes: negative, neutral and positive. We classify a

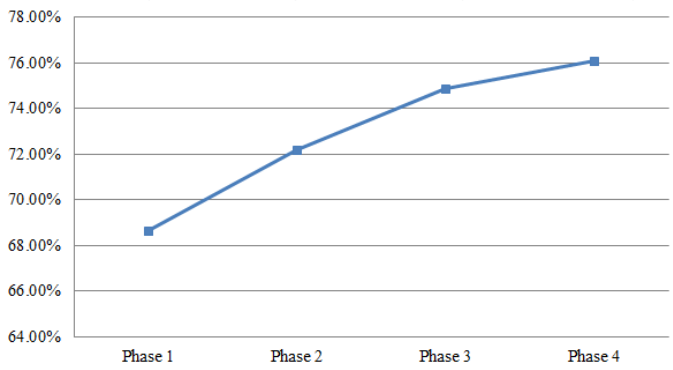

Fig. 3. The 10-fold cross validation accuracy on negative and positive tweets after each phase.

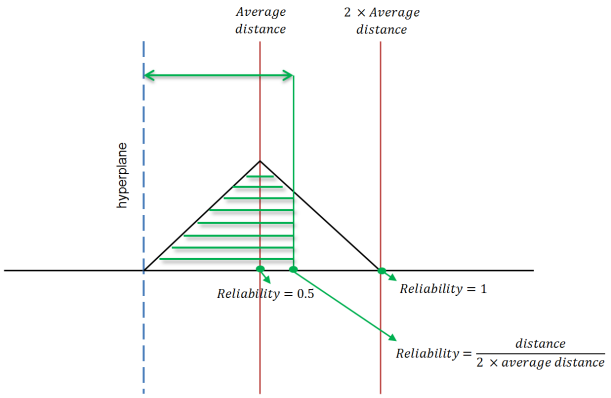

Fig. 4. Reliability as a function of distance from the SVM hyperplane.

tweet as neutral if the reliability of the binary classification is below a predefined threshold. The idea is based on geometric interpretation of the results provided by linear classifiers, and estimation of reliability from the provided scores [14].

A binary SVM classifier classifies an unseen instance as positive or negative, depending on which side of the hyperplane it falls. By taking the distance from the hyperplane into account, we can estimate the reliability of classification. Let d_A be the average distance of training examples (positive or negative) from the SVM hyperplane. Let d be the distance of the instance to be classified from the SVM hyperplane. The reliability R of the new instance classification is defined as

$$R = \begin{cases} \frac{d}{2 \times d_A} & \text{if } d < 2 \times d_A \\ 1 & \text{if } d \geq 2 \times d_A \end{cases}$$

When the classified instance is 'far enough' (more than two average distances) from the hyperplane, we consider it to be classified with reliability $R = 1$. Instances on the hyperplane have $R = 0$. Reliability therefore ranges between 0 and 1, where the higher values correspond to more confident classifications (see Figure 4). The definition is based on the assumption that for the whole population, distances from the SVM hyperplane are normally distributed around the average distance of the training examples. Area under the normal Gaussian curve is approximated by the area under the triangle, as illustrated in Figure 4.

We conducted a series of experiments by varying the reliability threshold R and testing by 10-fold cross validation. For these experiments we used the complete annotated dataset without the excluded tweets, i.e., the 26,175 general Bulgarian tweets. Only negative and positive tweets were used for training, but neutral tweets were predicted when they have reliability below the threshold. The threshold R was varied from 0 to 0.5, with an increment of 0.05.

With the three class classification, we are interested in the separation of positive from the union of negative and neutral tweets, and negative from the union of positive and neutral tweets. We plot ROC curves for both settings to get an insight into which reliability threshold values are optimal. A ROC curve [12] indicates the performance of a binary classifier system when its distinctive threshold is varied. It plots the fraction of true positives out of the positives (true positive rate) on the Y axis and the fraction of false positives out of the negatives (false positive rate) on the X axis, at different threshold settings. It illustrates comparative trade-offs between benefits (true positives) and costs (false positives). The respective ROC graphs are in Figures 5 and 6. A diagonal from the bottom left to the top right corner represents a random classifier. Points above the diagonal are better than random, and points closest to the left upper corner are 'the best'.

Figures 5 and 6 give a rough insight into which reliability threshold to choose. For the Bulgarian elections, we have chosen $R = 0.2$ as it seems a fair compromise between the sensitivity and specificity of predictions. However, different thresholds could be used for different classes, or we could aim at such a neutral zone which yields the same proportion of neutral tweets as observed in the training data.

G. Evaluation on gold standard: the three class problem

One goal of this subsection is to evaluate the performance of the three class classifier after the introduction of the neutral zone. Another goal is to estimate if the cross validation results obtained from training and testing on the general Bulgarian tweets are also valid for the more specific, political tweets.

We have randomly selected a subset of 90 tweets from the political Twitter corpora: 30 classified as positive, 30 as negative, and 30 as neutral (the later are those with reliability below 0.2). The tweets were carefully inspected by a native speaker and manually labeled as positive, negative or neutral, thus producing the gold standard set for evaluation.

First, to get results comparable to the cross validation tests, the reliability threshold of the classifier is set to $R = 0$, i.e., all tweets are classified just as positive or negative. The confusion matrix is in Table I. The resulting accuracy on negative and positive tweets is 80.3%, comparable to the cross validation accuracies reported in Section II-E.

Second, we set the reliability threshold to $R = 0.2$, thus predicting the three classes. In this setting, the 3-class accuracy is 53.3% (note that the majority baseline 3-class accuracy is 37.4%). True positive rates for negative, neutral and positive class are 0.56, 0.38, and 0.63, respectively. The confusion matrix is in Table II.

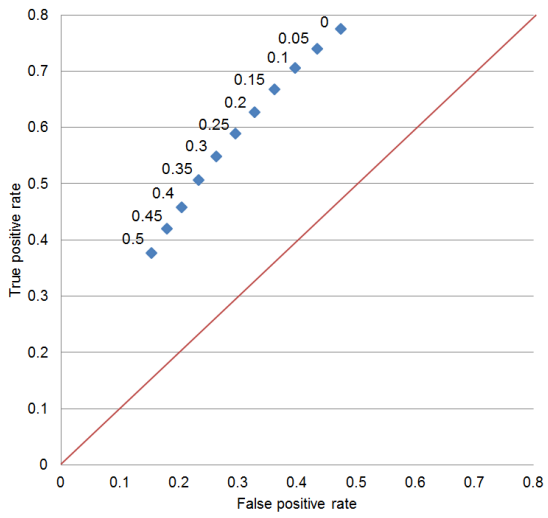

Fig. 5. The ROC points for "*positive* vs. *neutral-or-negative* tweets" by varying the reliability R from 0 to 0.5.

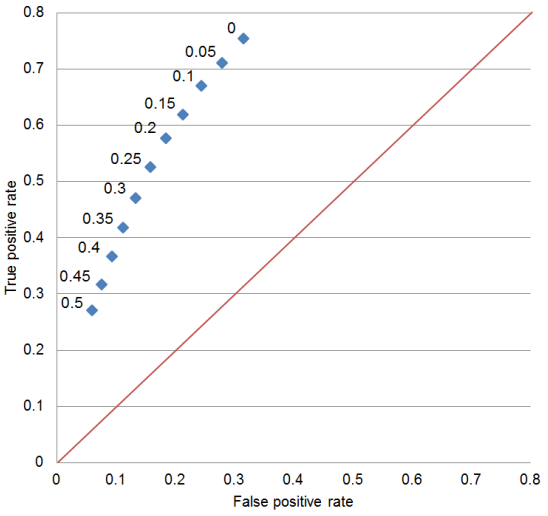

Fig. 6. The ROC points for "*negative* vs. *neutral-or-positive* tweets" by varying the reliability R from 0 to 0.5.

Careful inspection of the most obvious prediction errors revealed that the classifier could not recognize sarcasm (and therefore marked such tweets with incorrect sentiment) and that determining the sentiment was often difficult due to the unknown context (tweets are short and users tend to write compact messages). Moreover, we noticed that most of the features related to the names of the parties and politicians carry negative sentiment which might bias the sentiment negatively, despite the fact that the tweet contains positive or neutral opinion.

III. Analysis of the Elections Results

Parliamentary elections were held in Bulgaria on May 12, 2013. The 240 members of the parliament are elected by a proportional system, with a minimum of 4% of the votes to win a seat. In the analysis of the results, we consider only the following four political parties which passed the 4% threshold: GERB (Citizens for European Development of Bulgaria), BSP (Bulgarian Socialist Party), DPS (Movement for Rights and Freedoms), and ATAKA (Attack).

TABLE I. Confusion matrix on hand-labeled political tweets. Reliability threshold $R = 0$, i.e., there are no tweets predicted as neutral. The 2-class accuracy is 80.3%.

Predicted \ Actual	*Negative*	*Neutral*	*Positive*
Negative	31	13	5
Neutral	0	0	0
Positive	8	11	22

TABLE II. Confusion matrix on hand-labeled political tweets, reliability threshold $R = 0.2$. The 3-class accuracy is 53.3%.

Predicted \ Actual	*Negative*	*Neutral*	*Positive*
Negative	22	7	1
Neutral	12	9	9
Positive	5	8	17

The general mood of the voters before the elections was apathy, due to scandals and disappointment with politicians. Additionally, on the day before the election, when no political campaign is allowed, 350,000 alleged illegally printed ballots were discovered. All these contributed to the prevailing negative mood which was reflected in the Twitter sentiment.

We analyze 10,300 Bulgarian political tweets, fetched from Twitter between April 29, 2013 and May 15, 2013. We collected all the tweets which referred to the four largest political parties (which made it into the parliament) and political leaders of the corresponding parties.

We used a sentiment classifier constructed after the phase 4 of the annotation process. The training set consisted of negative and positive tweets only, ignoring the neutral tweets. The preprocessing settings which maximize the accuracy on the training set were used (see Subsection II-E). The only exception is that all punctuation symbols were removed. Reliability R was set to 0.2, i.e., all the tweets with reliability below 0.2 were classified as neutral.

The annotated tweets used for training were *general* Bulgarian tweets, the trained classifier, however, was applied to the *political* tweets. The sentiment distribution of both sets is in Figure 8, obviously significantly different.

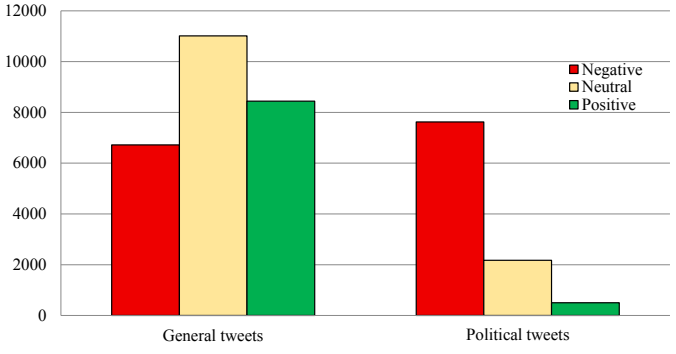

Fig. 8. Distributions of negative, neutral and positive tweets in the training and application datasets. At left: 26,175 general tweets, at right: 10,300 political tweets.

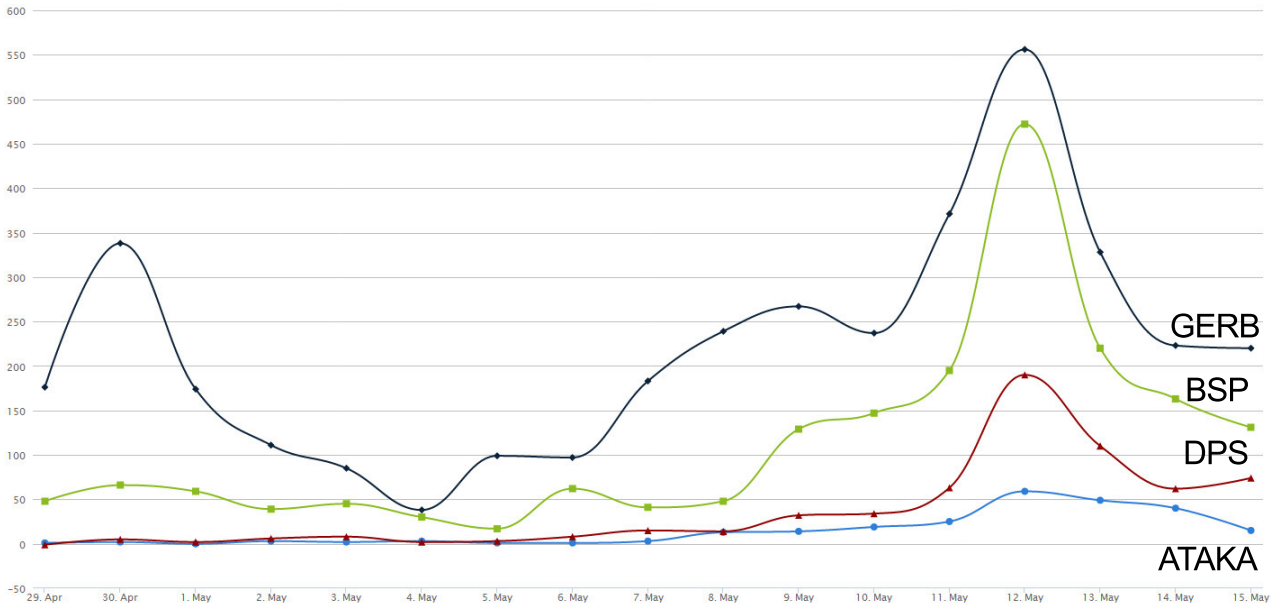

Fig. 7. The number of the *negative* minus *positive* tweets about the four political parties, before and after the Bulgarian elections (held on May 12, 2013).

Figure 7 shows the sentiment graph for the four parties in the monitored period. The graph reflects both the volume of tweets and the sentiment, since the y axis corresponds to the difference, the number of negative minus the number of positive tweets. We observe a prevailing negative sentiment and a large peak of activity on the elections day, May 12. In the following subsections we split the analysis of political tweets in two periods: before the elections (April 29 to May 11), and after the elections (May 12 to May 15). Tweets of May 12 were assigned to the post-election period because we observed that already early in the day some preliminary elections results (or exit polls) were published.

A. *Pre-elections analysis*

In this subsection we analyze Bulgarian political tweets which were posted before the parliamentary elections, i.e., between April 29 and May 11, 2013.

TABLE III. THE NUMBER OF NEGATIVE, NEUTRAL AND POSITIVE TWEETS PER PARTY, BEFORE THE ELECTIONS.

Party	*Negative*	*Neutral*	*Positive*	Total volume
GERB	2743 (74.4%)	771 (20.9%)	174 (4.7%)	3688 (100%)
BSP	1087 (74.8%)	284 (19.6%)	82 (5.6%)	1453 (100%)
DPS	244 (51.6%)	158 (33.4%)	71 (15.0%)	473 (100%)
ATAKA	123 (60.6%)	62 (30.5%)	18 (8.9%)	203 (100%)

TABLE IV. ELECTION RESULTS, VOLUME OF TWEETS, AND DIFFERENCE BETWEEN THE NEGATIVE AND POSITIVE TWEETS PER PARTY, BEFORE THE ELECTIONS.

Party	Parliamentary seats	Total volume	*Negative−Positive*
GERB	97 (40.4%)	3688 (63.4%)	2569 (66.7%)
BSP	84 (35.0%)	1453 (25.0%)	1005 (26.1%)
DPS	36 (15.0%)	473 (8.1%)	173 (4.5%)
ATAKA	23 (9.6%)	203 (3.5%)	105 (2.7%)
Total	240 (100%)	5817 (100%)	3852 (100%)

The number of tweets for each party is in Table III. It seems that before the elections, negative sentiment was prevailing, since for the major two parties the percentage of negative tweets is almost 75%, and only tweets about DPS are slightly less negative, but still over 50%.

Table IV shows the actual election results (in terms of the proportion of parliamentary seats won by each of the parties) in comparison to the volume of tweets, and differences between the negative and positive tweets. Most of the tweets (63% of the volume) before the elections mention the GERB party or its leader, with an even higher proportion of the negative−positive tweets (67%). GERB was eventually a relative winner of the elections, with the largest number of parliamentary seats (40%). Furthermore, the ranking of the major parties, both in terms of the volume and the negative−positive differences, corresponds to the actual elections results.

B. *Post-elections analysis*

In this subsection we analyze Bulgarian political tweets which were posted between May 12 and May 15, 2013, on the elections day and a few days after the elections. The analysis is similar to the previous subsection.

Table V gives the number tweets for each party. The negative sentiment is even higher than in the pre-elections period, possibly due to the ballot scandal on the day before the elections.

Table VI shows the number and proportion of tweets per party for the post-elections period. Again, most of the tweets (43% of the volume) refer to the GERB party or its leaders. This fraction, however, is lower than before the elections (63%) which indicates that after the elections the political discussions are more evenly spread over all the parties. Furthermore, the ranking of the elections results and even the proportion of the parliamentary sits won closely correspond to both, the tweets volume and the negative−positive sentiment differences. We compute the mean absolute error (MAE, an average of the absolute errors) which measures how close the predictions (the tweets volume or sentiment) are to the actual elections results

TABLE V. THE NUMBER OF NEGATIVE, NEUTRAL AND POSITIVE TWEETS PER PARTY, AFTER THE ELECTIONS.

Party	*Negative*	*Neutral*	*Positive*	Total volume
GERB	1510 (78.2%)	367 (19.0%)	54 (2.8%)	1931 (100%)
BSP	1139 (79.4%)	262 (18.3%)	33 (2.3%)	1434 (100%)
DPS	522 (72.4%)	168 (23.3%)	31 (4.3%)	721 (100%)
ATAKA	255 (64.2%)	102 (25.7%)	40 (10.1%)	397 (100%)

TABLE VI. ELECTION RESULTS, VOLUME OF TWEETS, AND DIFFERENCE BETWEEN THE NEGATIVE AND POSITIVE TWEETS PER PARTY, AFTER THE ELECTIONS.

Party	Parliamentary seats	Total volume	*Negative−Positive*
GERB	97 (40.4%)	1931 (43.1%)	1456 (44.6%)
BSP	84 (35.0%)	1434 (32.0%)	1106 (33.8%)
DPS	36 (15.0%)	721 (16.1%)	491 (15.0%)
ATAKA	23 (9.6%)	397 (8.9%)	215 (6.6%)
Total	240 (100%)	4483 (100%)	3268 (100%)

(the proportion of the parliamentary seats won). The MAE for the tweet volume is 1.88%, and for the negative−positive sentiment differences is 2.09%. In comparison, MAE for professional polling services is usually about 2-3% [15]. This leads to the conclusion that both closely correspond to the elections results, with the volume correlating slightly higher. As far as the sentiment is concerned, this is not surprising, since most of the tweets were negative.

IV. LANGUAGE DETECTION

This section describes the process of language identification of tweets. The Twitter API provides a language detection mechanism, but often it does not correctly distinguish between very similar languages, such as Bulgarian and Macedonian in our case. We describe two approaches to language detection, both based on building a classifier. We evaluate them by 10-fold cross validation and on gold standard Twitter data. The selected approach, actually used in our experiments, has the Bulgarian language detection accuracy of over 96%.

A. Problem

The majority of text mining and natural language processing tools are language-specific. Sentiment analysis tools are no exception as similar or exact same words in different languages may carry different sentiments (e.g. the word *gift*, which in English means a present, and a poison or toxin in German, can be associated with a positive or negative sentiment in the respective language).

Separate language detection models are required for each language in order to perform sentiment analysis in a multilingual setting. Language detection is performed when applying the sentiment analysis model, and also when building a model, since that requires manual annotation of tweets (described in section II-B), and the tweets are required to be in a language that is understood by the annotators.

For building and applying the models for the Bulgarian elections, we focused on the Bulgarian language. The Twitter API used for the collection of tweets provides meta-data on each tweet, which also includes the language detected by Twitter. At the time of the Bulgarian elections, this language detection was found to be far from perfect. The Twitter language detection could not correctly identify Macedonian tweets, and all Macedonian tweets were falsely labeled as Bulgarian (as the two languages are closely related). We implemented and trained a classifier for distinguishing Bulgarian from Macedonian.

B. Existing approaches

Language classification methods rely on statistical properties of text and supervised learning from given reference languages. To train a language detection classifier, the training data (a language corpus) is divided into smaller parts like letters, sequences of letters or whole words. Two popular methods for building language classification exist which differ in the strategy employed when dividing text into parts.

The *word-based method* tokenizes the text into words. Some word-based methods build discriminators for languages by using all the words from corresponding dictionaries of languages. Other methods focus only on specific words, as it is not necessary to take into account all the words in a dictionary. One method uses only short frequent words of four to five letter maximum length [19], [21]. Another method generalizes it and takes a specific number of the most frequent words of arbitrary length [9], [40].

In contrast to the word-based methods, *n-grams* can be analyzed to build language-detection models [7]. N-grams are sequences of letters of length n created by slicing words. It has been found that n-gram methods offer better precision and reliability over word-based methods [7], while other research points out that standard tools are not sufficient for distinguishing languages with a large lexical overlap (such as Bulgarian and Macedonian) [43]. A popular implementation of the n-gram method is the Compact Language Detection of the Google Chrome browser (http://src.chromium.org/viewvc/chrome/trunk/src/third_party/cld/). This method tokenizes text into 4-grams and compares them against a large table of reference 4-grams that have language properties associated with them. The accuracy of this detector increases with the length of the text, but similar language pairs pose challenges to this implementation.

C. Bulgarian and Macedonian language corpora

In order to train a language classifier we obtained a parallel corpus of Bulgarian and Macedonian news articles published by SETimes from the OPUS website [42].

The corpus consists of approximately five million words for each language. The content of the corpora is identical and aligned, and thus eliminates any content bias for a language. The SETimes news website publishes news and views from Southeast Europe in nine languages: Bulgarian, Bosnian, Greek, English, Croatian, Macedonian, Romanian, Albanian and Serbian.

D. Naive Bayes classifier

We have implemented a multinomial naive Bayes classifier which captures word or n-gram frequency in documents [28]. By using the multinomial method we assume that the length of the document is independent of its class (i.e., the language). This was done due to learning languages from news articles and applying the models to tweets, which are generally much shorter. The initial distribution of languages in texts (50% for

TABLE VII. LANGUAGE DETECTION PERFORMANCE BASED ON DIFFERENT STRATEGIES FOR TOKENIZATION, DETERMINED BY 10-FOLD CROSS VALIDATION.

Tokenization	Accuracy	Precision	Recall
Word-based	**96.99%**	**96.99%**	**97.13%**
4-gram	94.03%	93.30%	94.94%
3-gram	94.44%	93.89%	95.19%
2-gram	93.51%	92.52%	94.68%

TABLE VIII. LANGUAGE DETECTION PERFORMANCE BASED ON DIFFERENT STRATEGIES FOR TOKENIZATION ON TWEETS EXTRACTED FROM THE TWITTER API BY USING GEOGRAPHICAL LOCATIONS AS QUERIES.

Tokenization	Accuracy	Precision	Recall
Word-based	**96.00%**	**96.92%**	**96.85%**
4-gram	90.04%	89.43%	91.17%
3-gram	89.66%	88.89%	90.28%
2-gram	87.60%	84.98%	89.67%

each language) has been disregarded since we cannot be sure what the ratio of Bulgarian to Macedonian tweets is for an arbitrary search query to the Twitter API.

In order to determine which works best for the problem at hand we implemented several approaches for the tokenization of text: word-based, 2-gram, 3-gram and 4-gram.

E. Evaluation

The trained models were evaluated using 10-fold cross validation. In order to test in a similar setting that the classifier will be used on, we split the documents into sentences and evaluated each sentence, since sentences are more appropriate representations of tweets than whole documents.

The results of the cross validation are in Table VII. The word-based approach outperforms the n-gram methods for discriminating short texts in Bulgarian and Macedonian.

To further evaluate the classifiers, we extracted tweets from the Twitter API using the geographical locations of the capital cities of Bulgaria and Macedonia as queries. We collected 40,000 tweets, which were evenly distributed between the two geographical locations. We labeled the tweets from Bulgaria as Bulgarian and the tweets from Macedonia as Macedonian. We applied the classifiers on this data and calculated the accuracy, precision and recall. The results are in Table VIII. As in the cross validation test on the training data, the word-based approach performed best.

This model was then used to remove Macedonian tweets that were falsely labeled as Bulgarian by the language detection mechanisms, as implemented by Twitter.

V. CONCLUSIONS

We present a novel approach to real-time monitoring of sentiment on Twitter during interesting events, based on a relatively straightforward application of several machine learning and text mining techniques. The novelty and relative advantage of our approach is the construction of high quality sentiment classifiers from the high quality annotated data. The price to be paid is a relatively costly manual annotation of a considerable number of tweets used for training. We applied text preprocessing, specialized for Twitter, to generate appropriate feature vectors from the tweets. We have extended the binary SVM classifier into a 3-class classifier by introducing the neutral zone, where one can vary the tradeoff between the sensitivity and specificity of predictions. We also present an improved Twitter language detection which can be a difficult problem for similar languages (e.g., distinguishing between Slovenian, Croatian, Bosnian and Serbian).

The main issue in this work was the treatment of *neutral* tweets. If annotators are careful and do not use the neutral label as a "catch-all" class, but clearly distinguish between the "not polarized" sentiment and the "inappropriate or irrelevant" tweets, then the neutral label carries very useful information, and should be taken into account. Another issue was a highly negative sentiment observed early in political tweets which starkly contrasted a normally balanced sentiment in the general Bulgarian tweets (see Figure 8). This lead us to focus primarily on the distinction between the negative and positive tweets, thus ignoring the neutral tweets when building a classification model.

In our subsequent work, we treat the neutral class on par with the negative and positive classes, and a binary classification problem becomes a 3-class problem. Further, we consider the classes ordered (*negative* $\prec$ *neutral* $\prec$ *positive*) and construct an ordinal regression classifier with two SVM hyperplanes. This approach was successfully applied to diverse domains, such as the study of emotional dynamics in Facebook comments [47], the effects of Twitter sentiment on stock prices [33], and the sentiment leaning of different network communities towards environmental topics [36]. Extensive evaluations indicate that the classifier performance reaches the agreement of human annotators when the number of annotated tweets is large enough.

There are several venues of further research. Inspection of individual tweets has reveled that the names of political parties are sentiment-bearing, and that these sentiments vary between the names of the parties. This could result in two different sentiment scores for a tweet with the same content but referring to two different parties. In order to ensure that the tweets are classified based on the content alone, regardless of the political party, entities of interest should be replaced by tokens that do not bear any sentiment.

Tweets in the neutral class are typically hard to classify and could be further explored. One can analyze them to determine whether they have a highly positive and negative sentiment at the same time, if they are sarcastic, or if they are actually non-opinionated.

ACKNOWLEDGMENTS

This work was funded in part by the EU projects MULTIPLEX no. 317532, SIMPOL no. 610704, and DOLFINS no. 640772, and by the Slovenian ARRS programme no. P2-0103. We thank Matevž Gačnik, Gama System (http://www.gama-system.si), who initiated the project and provided the PerceptionAnalytics platform; our Bulgarian partners, Marius Markov and Alexander Milanov, and the annotators for timely labeling Bulgarian tweets in the short period before the elections; Sowa Labs (http://www.sowalabs.com) for providing the Goldfinch annotation platform; Matjaž Juršič for help with DB issues; and Dragi Kocev for careful inspection and annotation of a sample of political tweets.

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment analysis of twitter data, In: Proc. Workshop on Languages in Social Media, pp. 30–38, 2011.

[2] S. Asur, B.A. Huberman, Predicting the future with social media, In: Proc. IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology, pp. 492–499, 2010.

[3] A. Bermingham, A.F. Smeaton, On using Twitter to monitor political sentiment and predict election results, In: Sentiment Analysis where AI meets Psychology, pp. 2-10, 2011.

[4] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science 2(1): 1–8, 2011.

[5] J. Borondo, A.J. Morales, J.C. Losada, R.M. Benito, Characterizing and modeling an electoral campaign in the context of twitter: 2011 spanish presidential election as a case study, Chaos: An Interdisciplinary Journal of Nonlinear Science 22(2), 2012.

[6] G. Caldarelli, A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, G. Riotta, A multi-level geographical study of Italian political elections from Twitter data, PloS One 9(5):e95809, 2014.

[7] W.B. Cavnar, J.M. Trenkle, N-Gram-Based Text Categorization, In: Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175, 1994.

[8] J. Chung, E. Mustafaraj, Can collective sentiment expressed on Twitter predict political elections, In: Proc. 25th AAAI Conf. on AI, San Francisco, USA, 2011.

[9] J. Cowie, Y. Ludovic, R. Zacharski, Language Recognition for Mono- and Multi-lingual Documents, In: Proc. Vextal Conference, Venice, 1999.

[10] J. DiGrazia, K. McKelvey, J. Bollen, F. Rojas, More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior, SSRN http://ssrn.com/abstract=2235423, 2013.

[11] Y-H. Eom, M. Puliga, J. Smailović, I. Mozetič, G. Caldarelli, Twitter-based analysis of the dynamics of collective attention to political parties, PLoS ONE 10(7): e0131184, 2015.

[12] T. Fawcett, An introduction to ROC analysis, Journal Pattern Recognition Letters 27(8): 861–874, 2006.

[13] R. Feldman, J. Sanger, The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007.

[14] P. Flach, Machine Learning, Cambridge University Press, 2012.

[15] D. Gayo-Avello, P.T. Metaxas, E. Mustafaraj, Limits of electoral predictions using Twitter, In: Proc. Intl. Conf. on Weblogs and Social Media, 2011.

[16] D. Gayo-Avello, Don't turn social media into another 'Literary Digest' poll, Communications of the ACM 54(10): 121–128, 2011.

[17] D. Gayo-Avello, "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"–A Balanced Survey on Election Prediction using Twitter Data, arXiv preprint http://arxiv.org/abs/1204.6441, 2012.

[18] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, 2009.

[19] G. Grefenstette, Comparing Two Language Identification Schemes, In: Proc. 3rd Intl. Conf. on the Statistical Analysis of Textual Data, Rome, 1995.

[20] R. Holt, Twitter in numbers, http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html, 2013.

[21] N.C. Ingle, A Language Identification Table, The Incorporated Linguist 15(4): 98-101, 1976.

[22] K. Jahanbakhsh, Y. Moon, The Predictive Power of Social Media: On the Predictability of US Presidential Elections using Twitter, arXiv preprint http://arxiv.org/abs/1407.0622, 2014.

[23] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent Twitter sentiment classification, In: Proc. 49th Annual Meeting of the Association for Computational Linguistics, pp. 151-160, 2011.

[24] T. Joachims, Making Large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning, MIT Press, 1999.

[25] A. Jungherr, P. Juergens, H. Schoen, Why the pirate party won the german election of 2009 or the trouble with predictions: A response to A. Tumasjan et al., Social Science Computer Review 30(2): 229–234, 2012.

[26] A. Jungherr, Tweets and votes, a special relationship: The 2009 federal election in Germany, In: Proc. 2nd Workshop on Politics, Elections and Data, ACM, pp. 5–14, 2013.

[27] A. Livne, M.P. Simmons, E. Adar, L.A. Adamic, The Party Is Over Here: Structure and Content in the 2010 Election, In: Proc. Intl. Conf. on Weblogs and Social Media (ICWSM), 2011.

[28] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, In: AAAI Workshop on learning for text categorization, AAAI Press, pp. 41-48, 1998.

[29] P.T. Metaxas, E. Mustafaraj, D. Gayo-Avello, How (not) to predict elections, In: Privacy, security, risk and trust (PASSAT) and IEEE 3rd Intl. Conf. on Social Computing (SocialCom), pp. 165–171, 2011.

[30] B. O'Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: Linking text sentiment to public opinion time series, In: Proc. Intl. Conf. on Weblogs and Social Media (ICWSM), pp. 122–129, 2010.

[31] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, In: Proc. ACL Conf. on Empirical Methods in Natural Language Processing, pp. 79–86, 2001.

[32] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2(1-2): 1-135, 2008.

[33] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, I. Mozetič, The Effects of Twitter Sentiment on Stock Price Returns, arXiv preprint http://arxiv.org/abs/1506.02431, 2015.

[34] E.T.K. Sang, J. Bos, Predicting the 2011 dutch senate election results with twitter, In: Proc. Workshop on Semantic Analysis in Social Media, ACL, pp. 53–60, 2012.

[35] M. Skoric, N. Poor, P. Achananuparp, E.P. Lim, J. Jiang, Tweets and votes: A study of the 2011 Singapore general election, In: 45th Hawaii Intl. Conf. on System Science (HICSS), IEEE, pp. 2583–2591, 2012.

[36] B. Sluban, J. Smailović, S. Battiston, I. Mozetič, Sentiment Leaning of Influential Communities in Social Networks, Computational Social Networks 2(9), 2015.

[37] J. Smailović, M. Grčar, M. Žnidaršič, N. Lavrač, Predictive sentiment analysis of tweets: A stock market application, Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Lecture Notes in Computer Science (Vol. 7947), pp. 77–88, 2013.

[38] J. Smailović, M. Grčar, N. Lavrač, M. Žnidaršič, Stream-based active learning for sentiment analysis in the financial domain, Information Sciences 285: 181–203, 2014.

[39] J. Smailović, Sentiment analysis in streams of microblogging posts, PhD Thesis, Jozef Stefan Intl. Postgraduate School, Slovenia, 2015.

[40] C. Souter, G. Churcher, J. Hayes, J. Hughes, Natural Language Identification Using Corpus-Based Models, Hermes Journal of Linguistics, pp. 183-203, 1994.

[41] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment in Twitter events, Journal of the American Society for Information Science and Technology 62(2): 406-418, 2011.

[42] J. Tiedemann, News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces, In: Recent Advances in Natural Language Processing (Vol. V), pp. 237-248, 2009.

[43] J. Tiedemann, N. Ljubesic, Efficient Discrimination Between Closely Related Languages, In: Proc. 24th Intl. Conf. on Computational Linguistics (COLING), Mumbai, India, 2012.

[44] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, In: Proc. Intl. Conf. on Weblogs and Social Media (ICWSM), pp. 178–185, 2010.

[45] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Where there is a sea there are pirates: Response to Jungherr et al., Social Science Computer Review 30(2): 235–239, 2012.

[46] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[47] F. Zollo, P.K. Novak, M. Del Vicario, A. Bessi, I. Mozetič, A. Scala, G. Caldarelli, W. Quattrociocchi, Emotional Dynamics in the Age of Misinformation, to appear in PLoS ONE, arXiv preprint http://arxiv.org/abs/1505.08001, 2015.