A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter

Lazaros Oikonomou
The Data Mining and Analytics Research Group,
School of Science & Technology,
International Hellenic University,
Thermi, Greece
l.oikonomou@ihu.edu.gr

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

***Abstract*— This paper presents work on using data extracted from Twitter to predict the outcome of the latest USA presidential elections on 8th of November 2016 in three key states: Florida, Ohio and N. Carolina, focusing on the two dominant candidates: Donald J. Trump and Hillary Clinton. Our method comprises two steps: pre-processing and analysis and it succeeded in capturing negative and positive sentiment towards these candidates, and predicted the winner in these States, who eventually won the presidency, when other similar attempts in the literature have failed. We discuss the strengths and weaknesses of our method proposing directions for further work.**

Keywords*— *Sentiment analysis; Classification; Social Media; Data mining; Artificial Intelligence and Applications

I. INTRODUCTION

Social media is a part of our lives for some years now. People increasingly tend to express their opinions via social platforms. On a daily basis, data that originate from social media are massive in terms of volume. The question this paper addresses is the following: Can we use these data in order to detect trends, preferences, patterns and therefore predict outcomes of future events?

Social media, and more specifically Twitter, research has become more and more intense over the last years. Twitter is considered one of the most successful social media of our time. The community of the popular platform counts more than 328.000.000 active users at the moment [1]. As a result, the volumes of data created on a daily basis are vast. The platform offers the opportunity to its members to write text messages – or as the Community calls them, tweets – to address any topic they want. A lot of researchers have tried to evaluate Twitter data to predict future outcomes. The concept behind all studies in the field, is to collect tweets using Twitter's API and apply different algorithms in order to classify them and find trends in what users are saying about a specific topic. This can help data analysts foresee outcomes and make low risk assumptions about various cases.

Despite the number of studies that appertain to Twitter and its data, there are challenges that still exist, such as: the optimal way to gather data, since Twitter offers a variety of choices regarding API calls and the method that data will be analyzed in order to achieve high accuracy for predictions in the final results. The focus of this study is to face these challenges and arise solutions that are credible and concrete.

In particular, this paper focuses on the presidential elections in USA that were scheduled for November 8th 2016. The goal was to gather tweets that refer to the elections and more specifically to the two main candidates: Hillary Clinton and Donald J. Trump. After acquiring our data, the method proposed consists of a classification algorithm selection and implementation. In order to attain classification on text, the key term is 'sentiment analysis' (or opinion mining). Under this perspective, in the end of this paper we present our prediction about the election results based on the method proposed.

There are two main challenges for this work:

- Data gathering. This task included a series of decisions that needed to be made. The Twitter Developer platform is a powerful tool that offers different approaches according to the goals and demands of each project. Calls to the Twitter API are free, upon request, and can be addressed using two different approaches:

  1. The Streaming API
  2. The Search API

  After deciding which API was the optimum for this research project's needs, there were some other challenges that had to be addressed. The tweets had to be originated from US citizens located within the USA region. People who would eventually vote at the elections were the study's target. Gathering tweets from all the states was not mandatory because in some States the polls were already clear on who would win. In eleven states the polls were indecisive among the two dominant candidates. The States chosen to retrieve data from were the most controversial and historically decisive [2], [3].

- Sentiment analysis. The tweets retrieved after the data gathering process were plain text, a combination of subjective and objective words. In order to get the sentiment behind a

tweet, we gathered the subjective words from the text and checked if they had a negative or a positive impact on the opinion the user was trying to express. In order to classify our data, the Naïve Bayes classification algorithm was used. We chose this approach to achieve simplicity and high accuracy as we will explain in the following sections.

In the next sections we elaborate on our approach as follows: section II explores the background of the topic and outlines some key examples of related word. Section III, explains our proposed method. Experimental results are presented in section IV which compares our predictions with the actual results and discusses key findings. A discussion of our results along with a review of decisions made in order to optimally implement our plan is included in section V. Finally, the paper ends with conclusions and suggestions for further work in section VI.

II. BACKGROUND & LITERATURE REVIEW

Even though Social media analysis is a field that started being popular a few years back, there are a lot of researchers who tried to develop methods that aim to find what users believe about various topics. In our case, we review seminal works that targeted elections or candidates that participated in elections.

In 2012 before the French presidential elections Wolska and Bouguera published a paper called "tweets mining for French presidential election" [4]. Its purpose was not only to predict the election result, but also to analyze how different trends influence the masses in social media using different polling methods. Furthermore, the meaning of opinion intensity was introduced, giving the opportunity to draw conclusions about how loyal and determined voters are. As the authors claim, social network analysis and text mining for political purposes is a field that poses a lot of challenges and it might become a useful and accurate method of predicting both political and economic trends in the future.

In 2013 Mahmood et al. [5], published a paper on "Mining Twitter Big Data to Predict 2013 Pakistan Election winner". Their goal was to retrieve tweets, pre-process them, store them in a database and come up with ways to draw conclusions about the winner of the elections. Three parties were massively twitted at that time, Pakistan Tehreek-e-Insaaf (PTI), Pakistan Muslim League Nawaz (PMLN) and Muttahida Qaumi (MQM). PMLN was the party that prevailed at the end. The authors used three different approaches to complete their task. The three models were CHAID decision tree, Naïve Bayes model and Support Vector Machines (SVM). Their analysis showed that PTI would win, but the winner was different. However, there was in fact a twitter-based trend towards PTI that helped the party gain more votes than it was originally meant to get. This means that the party focused on social media promotion, to attract young voters. Consequently, the party was regularly mentioned in tweets giving the impression to users that it was higher on polls than it really was. This lead to a higher number of votes.

Soler et al. [6], published a paper in 2012 on "Twitter as a tool for predicting election results". This paper presented a new tool for Twitter analysis called TaraTweet. Using this tool 500.000 tweets were analyzed and the results were shown at TaraTweet's site. The results of all three experiments conducted had adequate accuracy. There are some parties that received different percentages of votes than originally predicted, which is normal, since voters may make their mind up until the very last minute. With these three experiments, we can identify a correlation between mentions and actual vote intentions. Parties that invest in social media promotion are more likely to see satisfying results in the elections. This paper indicates that Twitter analysis is a safe method to conduct experiments and come up with results that are really close to real-time vote intentions.

Jose and Chooralil [7] published a paper in 2015 that proposed a new approach for sentiment analysis. They used a novel method for analyzing tweets and with the help of lexical resources such as SentiWordNet, WordNet and word sense disambiguation they tried to extract information and knowledge out of tweets. In addition, in order to achieve the highest accuracy possible they also proposed a negation handling approach in the data pre-processing stage. The authors' innovative spirit lead them to try a variety of tools. The data gathering process was implemented with the help of Twitter's Streaming API.

Another interesting work by Akshi Kumar et al. [8], proposed a method for Emotion analysis of Twitter using Opinion mining called "Emotion analysis of Twitter using Opinion mining". The tweets were analyzed using sentiment analysis. However, they used a slightly different approach. According to the authors the basic sentiments that a person might have are: Happiness, Anger, Sadness, Fear and Disgust. These are the sentiments that Paul Ekman and his team found in their research in 1972 to be the most common among human beings [9]. The data that were gathered in this work were classified based on Ekman's approach. This approach may find several different uses like recommendation systems in business intelligence and to reveal vote intentions of social media users.

In 2014 Das et al. published a paper about sentiment analysis for products using public tweets from customers [10]. This paper took the topic one step further by building an application that helps researchers and business analysts to see how consumers feel about a released product. The system they proposed involved a tool for conducting research about different topics (new products, opinion about election candidates etc.). What it does, is data gathering from Twitter's streaming API and preprocessing in a way that sentiment analysis is easily done. The reports that were generated after the sentiment analysis of tweets is what matters most, because with their help, high-end users (like managers, political analysts etc.) could draw conclusions and foresee future events.

Tumasjan et al. tried to answer three questions regarding politics and Twitter: a) does Twitter provide a platform for political deliberation online? b) how accurately can Twitter inform us about the electorate's political sentiment? and c) can Twitter serve as a predictor of the election result? [11]. They examined 104,003 tweets related to the German elections (containing the names of the parties that participated). To extract their sentiment, they used LIWC2007 (Linguistic Inquiry and Word Count;). Their focus was on 12 dimensions:

future orientation, past orientation, positive emotions, negative emotions, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money. They concluded that Twitter is indeed a platform where users tend to express themselves politically, engaging with others and exchanging opinions. Regarding the second question, voters showed their sentiments towards the elections depending on their political leader views. Tweets about Steinmeier were more ambiguous due to the fact that the leader himself did not have a clear view on potential political coalitions. Finally, the results of the sentiment analysis, even though the sample of tweets was not large enough, were really close to the actual election results. In their opinion, Twitter can in fact be considered as a valid indicator of electorate's opinions.

Jungherr et al. focused on evidence that the scientific community was a victim of a common misconception: trying to draw conclusions based on social media metrics sometimes leads to false results due to the lack of thorough testing [12]. Their main concern was that there are a lot of political parties or other enterprises that involve "actors" that try to influence people with various means like propaganda and fake news. To prove their point, they collected data from a social media vendor called Gnip (http://gnip.com). Messages about the 2013 elections in Germany were retrieved and the sample contained 6,677,795 messages posted by 1,248,667 users. During their analysis they established a custom solution by hand-coding 1% of all political mentions in messages in order to identify positive, negative or neutral sentiment. Furthermore, they used Hopkins/King's approach to automate content analysis, a method that has been used in other similar projects as well. Finally they also used hashtags that included specific words like the parties' names. The results from their research indicate that the overall validity of Twitter-based mentions as indicators of political support as expressed in votes is rather poor. However, they suggested that Twitter is potentially a tool to measure voter sentiment. They also suggested that future research may focus on using Twitter data to analyze which kind of political information attracts Twitter users' attention and is distributed online and also be well advised to focus on theoretical underpinnings, rather than exclusively on empirics. Their large data collection and thorough analysis assert this work as a considerable contradiction to ours.

These are the 8 most prominent works in the field. There are other works that were also successful. The number of people trying to analyze data extracted from social media shows that this field is emerging. Every work included interesting perspectives and most of them were successfully implemented. In the following section, we will demonstrate how our work differentiates from other similar works. We will describe our method and the decisions we made in order to achieve high accuracy.

III. PROPOSED METHOD

The approach we followed comprises two main features

- Data gathering
- Sentiment analysis

A. Data gathering

The data gathering process is challenging. The dataset collected, has to be credible and preprocessed properly for the sake of integrity and accuracy of the final results. The challenges that this stage included were the following:

- API selection
- Constructing the right queries
- Data preprocessing.

1 API selection

The Twitter developer section has a detailed documentation for API calls which is very simple to understand. When someone wants to make requests to the Twitter API, an access token is required. To get the access token and thus the authorization to make requests to the API, Twitter requires from the user to build an application that will authenticate the requests. After building the application, the user is ready to make queries and gather data. The application built for this work is titled TweetGrabber13. The process to design and implement this application was simple as there are precise guidelines from Twitter explaining step by step how this can be done [13].

After building our application, the next step was to choose which API to use. The two APIs Twitter offers, have some similarities but they are also different in many ways. Both need OAuth, which is a specific type of user authentication in order to connect to Twitter. In addition, the data formats that both APIs return are the same (JSON).

One of the most important differences between the APIs is the rate limit each one offers. The Streaming API offers a much narrower rate limit than the Search API. This is one of the reasons that in this work, data were gathered using the Search API. We had to make requests during different periods of time and the Search API is the best choice for this kind of research. Another difference between the APIs is that the Streaming API provides real-time data, when the Rest API returns fewer results for every request submitted. When a query is built and executed with the Streaming API, all the tweets that match the criteria are retrieved. On the other hand, the Search API returns a sample of tweets that are more relevant to the query. Our target was to keep our dataset clear of irrelevant tweets, so once again the Search API was considered more suitable. Finally, the options that the Search API offers to build queries was also an asset for our decision. The parameters offered are numerous, helping data analysts build queries based on location, popularity, language and many more [13].

2 Querying with PHP

A programming language very suitable to make API calls is PHP. The variety of libraries, selectors and attributes PHP offers makes it one of the easiest programming language for communication with APIs [14].

After building the application we needed to choose the three states that the tweets would be collected from. The three states with the most controversial and ambiguous polls were:

- Florida

- Ohio
- North Carolina

These were the three states expected to determine the final results [2], [3].

Another important factor while collecting data was the various parameters that needed to be added to the queries. Tweets that referred to the candidates, the elections or any other relevant topic would be the target. To achieve getting the required results the queries were based on hashtags that invoked relevance to the topics addressed. Since the aim was to get opinions and statements about the two main candidates we chose hashtags that included the respective candidate campaigns. The tweets gathered included the following hashtags:

- #donaldtrump, tweets that are about Donald J. Trump
- #hilllaryclinton, tweets that are about Hillary Clinton
- #maga, official hashtag in favor of Donald J. Trump meaning (initials mean make America great again)
- #imwithher, official hashtag in favor of Hillary Clinton
- #nevertrump, official hashtag against Donald J. Trump's campaign
- #neverhillary, official hashtag against Hillary Clinton's campaign

Some of the tweets gathered are considered polarized beforehand. Tweets containing the hashtags "#maga", "#imwithher" are the official hashtags in favor of Donald Trump and Hillary Clinton campaigns. Those tweets 277.509 tweets were gathered in total. The data gathering process lasted one month. There requests were issued daily from October 7th to November 7th. The elections were scheduled for November 8th 2016.

3 Data preprocessing

In order to make our data easier to interpret, we preprocessed them to get rid of unnecessary information. The Search API returns a massive volume of metadata for every tweet. The fields that we were interested in were the following:

- Tweet's id
- Date posted
- Time posted
- Tweet's text
- User name
- User id
- Tweet's status (in order to check if it is a retweet or an original message)

These were the most important information we needed for each tweet. Furthermore, we decided to change the format of the data, transforming it in text files (txt). That decision was taken in order to make the results more readable and simpler to navigate.

B. Sentiment analysis

It is clear that classification of tweets is this work's target. Tweets had to be classified as positive, neutral or negative. One of the efficient ways to perform classification is the Naïve Bayes Model, known to be simple but yet powerful for classification [13]. This is the model we chose to classify our data with. A powerful tool that was also used is Textblob, a Python library that besides classification helped us set subjectivity scores to tweets. Below we elaborate on why we used these two options.

1 Naive Bayes Method

The Naïve Bayes classifier is based on Bayes' Theorem which defines the probability of an occurrence by observing prior data that are relevant to the event. For example, the probability of rain is higher when there are clouds in the sky. Thus, it is easier to predict the probability of raining when we know that the sky is full of clouds than if we did not. The implementation of the algorithm is simple code-wise and its performance surpasses similar approaches [15].

A classifier has to be trained in order to implement Naïve Bayes Theorem. A classifier is defined as a set of data that are already classified and help us construct rules that apply to our dataset as well. In order to train this classifier, a training dataset is required. There are many free available corpuses that can be downloaded and be used as training data. The corpus chosen for this work is a combination of the N. Sanders dataset and the corpus that Michigan University developed during a sentiment analysis competition on Kaggle [16]. Our training dataset consisted of 1.578.627 tweets that have already been classified as positive, neutral or negative. The sample is quiet large and provides a variety of words and phrases that can help create rules applicable to our experiment. The accuracy of this classifier was over 75%, a number that is above average, but could be further improved by experimenting with different algorithms. For accuracy purposes, we decided to also run our data through Textblob as we explain below.

The next step was to train our classifier based on the training corpus. The classifier followed a basic principle, which is: Tweets with polarity value less than 0 are negative, equal to 0 neutral and greater than 0 positive. The polarity score was assigned depending on how many positive or negative words are contained in it. We developed the respective application in Python. After the sentiment analysis stage was over, all tweets were classified. The process we followed and the answers to the key challenges we met are summarized in Fig. 1.

2 Textblob library

A powerful tool that was also used is Textblob. Textblob is a Python library we used to analyze textual data. It provides an API that can help implementation of different Natural Language Processing (NLP) tasks, one of which is sentiment and subjectivity analysis [17]. Subjectivity, is defined as the number of subjective words per sentence. A tweet that has a lot of words that express sentiment like adjectives, adverbs, verbs etc. has a higher subjectivity score than a tweet that does not. Subjectivity is an important factor and that is why we decided to also analyze our tweets with Textblob. Once we imported Textblob, we also imported some textblob.classifiers modules.

With these modules we can easily build customized classifiers like sentiment analyzers and more. Textblob is a powerful library and can be used in various data mining projects. There are 6 modules that we built in order to perform sentiment analysis on our data. Our code is based on Naïve Bayes classification model and what it does is to classify each tweet as positive, negative or neutral depending on their textual context and also set a subjectivity score for each tweet. So we define a polarity and subjectivity score for each tweet. There are three possible values for polarity: negative, positive and neutral. Our code sets a value between -1 and 1 for each tweet (-1 <= *polarity score* <= 1). If the polarity score is 0 the tweet is considered neutral, if the polarity score is less than 0 the tweet is considered negative and if the polarity score is greater than 0 then the tweet is listed as positive. The subjectivity score for each tweet ranges between 0 and 1 (0 <= *subjectivity score* <= 1). Now let's see how the modules we have on our code do what we described.

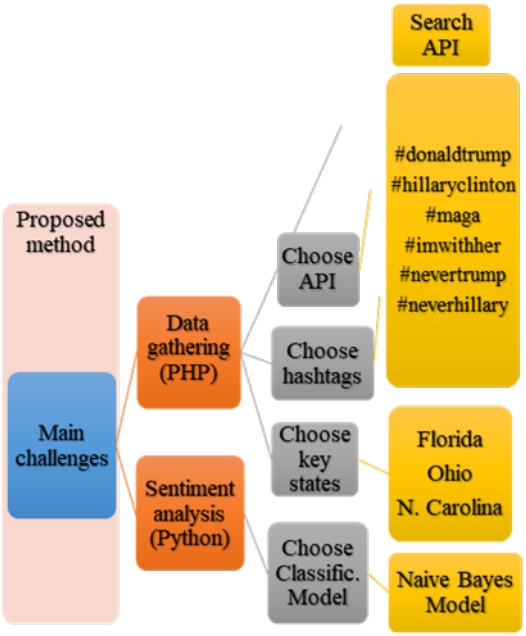

Fig. 1. Key challenges for the proposed method.

IV. EXPERIMENTAL RESULTS

This work aimed at predicting the USA elections in three key states that could potentially determine the final result. We will here present our sentiment analysis predictions, and how close these predictions were to the actual election results. These results were posted online the day before the elections [22].

A. Florida results

73263 tweets were gathered from the state of Florida. From these 73263 tweets 42683 were in favor of Donald Trump, 25938 were in favor of Hillary Clinton and 4642 tweets were classified as neutral. To sum up the results that our analysis predicted were:

- Donald Trump: 58.26%
- Hillary Clinton: 35.4%
- Neutral tweets: 6.44%

The actual results of the elections in the state of Florida were:

- Donald Trump: 49.1%
- Hillary Clinton: 47.8%
- Other Candidates: 3.1%

We can see that the percentages are fairly close, but the most important thing is that we correctly predicted who would win the state.

B. Ohio results

116733 tweets were gathered from the state of Ohio. From these tweets 65.611 tweets were in favor of Donald Trump, 41727 tweets were in favor of Hillary Clinton and 9395 were classified as neutral. So, the percentages that the analysis gave us were:

- Donald Trump: 56.2%
- Hillary Clinton: 35.75%
- Neutral tweets: 8.05%

Below we can see what the actual results were:

- Donald Trump: 52.1%
- Hillary Clinton: 43.5%
- Other Candidates: 4,6%

Like Florida, we can observe that the results are close, but the important thing is that once again we predicted who the final winner was.

C. N. Carolina results

The final state we gathered tweets from is N. Carolina. 87513 tweets were gathered in total from this state. From these 87513 tweets, 46608 were in favor of Donald Trump, 33301 were in favor of Hillary Clinton and 7604 tweets were classified as neutral. So, the percentages that our analysis predicted were:

- Donald Trump: 53.26%
- Hillary Clinton: 38.05%
- Neutral tweets: 8.69%

The actual results of the state were:

- Donald Trump: 50.5%
- Hillary Clinton: 46.7%
- Other Candidates: 2.8%

Once again, the results were really close to our predictions and the most important thing is that we correctly predicted who would win the state. In Fig. 2 we provide two charts with all our predictions compared to the final results of the elections. As it can be seen, we correctly predicted the winner in all three states.

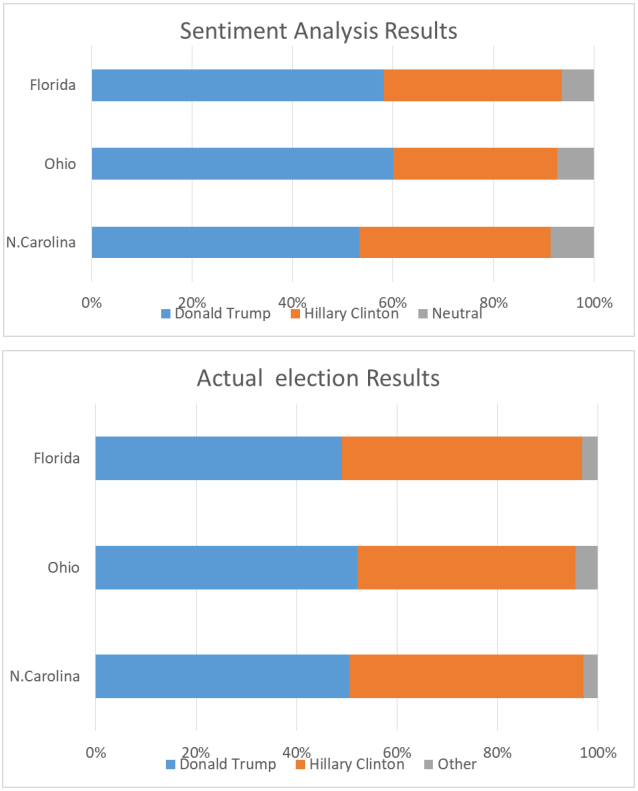

Fig. 2. Comparison of predictions [22] with actual results

In order to highlight the difficulty of the prediction and the significance of our prediction we cite some of the polls that well-known websites and newspapers published before the elections, predicting that Hillary Clinton would eventually win the race.

Fig. 3: Huffington post predictions towards the USA 2016 presidential elections [16].

Fig. 4: New York Times prediction for the USA 2016 presidential elections [19].

Fig. 5: CNBC posts UK betting firms predictions about the elections [20].

DISCUSSION

The importance of social media analysis in elections can be seen, once we consider the predictions that popular and credible media made. A recent study based on a paper written by D. Kreiss and S.C. Macgregor, highlights the importance of social media in the elections [23]. According to this work, the basic difference between Donald Trump and Hillary Clinton campaigns is that the first allowed employees of the top technological firms like Facebook, Twitter and Google to be part of his campaign and more specifically in the decision making of his candidacy, while the latter decided not to do so. This helped Trump do more targeted advertisements and reach people based on their individuality, based on the knowledge and abilities of the employees of the tech firms that knew precisely how to get in touch, digitally, with the voters. This may arouse questions regarding the credibility of the elections, but also highlights the importance of social media in events like elections.

The challenges for predicting elections using social media have been highlighted in a number of studies. In [24] the authors used a model for utilizing Twitter to predict the outcome of the UK general election. They projected a

parliament with 285 Conservative party seats and 306 Labour Party seats. Their prediction proved inaccurate as the Conservatives won the election with 330 seats to 232 by the Labour party.

In [25] the authors used 13 different variables that were available online including Tweets, Celebrity Tweets and Celebrity Sentiments, Twitter Followers, Facebook Page Likes and Wikipedia Traffic for the 2016 US Presidential Elections. Wikipedia page traffic for the Democratic Party candidates did not correlate with the outcome of the DP election since Sanders lost to Clinton even though his Wikipedia page traffic was at least twice larger than Clinton's for a period of 2.5 months from November 2015 up to mid-January 2016. The research found correlations between polls and Facebook page likes, and between polls and Twitter.

In [26] a research about whether Social Media can predict election results in New Zealand demonstrated that the Number of Facebook friends and the Twitter followers as explanatory variables are not good indicators as only 16.7 and 5.4 percent of election winners were predicted correctly respectively.

Our work was based on Twitter users that wanted to express their sentiment about two candidates. The analysis we made revealed an advantage for Donald Trump even though all polls showed the opposite. We can say that voters prefer to express their feelings via social media than a poll that in most cases is conducted through a phone call. People feel that their social media profile identifies them uniquely and that is why analysts must see social media as a trustworthy tool for their future works.

On the other hand, there are a few risks that we make when we draw conclusions from social media. There are cases of users that might try to compromise work like ours. Since political analysts realized the importance of Social media in election races, some of them tried to manipulate the data that exist on them. A lot of parties hire individuals that post positive comments about the party continuously. This is an issue that needs to be addressed and it is extremely important to be dealt with.

Another risk that needs our attention is the misclassification of tweets. There might be cases that some of the tweets are classified incorrectly. To make the final results accurate, the method developed must be well designed and ready to deal with any type of threat. In our work the percentage of correctly classified tweets was 95.5%.

V. CONCLUSIONS & FUTURE WORK

A. Conclusions

Having reviewed the results produced by our proposed method and our predictions for the elections that occurred on November 8th 2016, we can conclude that our experiment was successful. We predicted correctly who would win in all three key swing states we gathered tweets from. Consequently, we can claim that Twitter is a social media that analysts can use in order to draw useful conclusions about elections, but also possibly for other things as well, like products, events, market issues and so on.

Of course, we cannot claim that analyzing tweets is the ideal way of predicting election results with high accuracy, since a lot of voters do not have a Twitter account, but we can see what the public opinion says about different topics and make some low risk assumptions about the future. In this work, we tried to observe how the voters felt about the candidates, we did not count how many votes each one would get. Thus, we cannot directly predict the winner. However, the statistical analysis and polls that we cited in the previous section show that predicting the winner of such an event is not a simple task. Such problems should be addressed in multidimensional ways and take many factors into consideration before making a prediction. It should be noted that comparing the number of tweets with the number of votes does not imply a statistical significant correlation between the two variables.

In our case, what was really odd is that even though all polls showed that Hillary Clinton would eventually win the elections, Twitter users mainly mentioned Donald Trump when they posted a tweet about the elections. As a result, when we made a request with the Search API to retrieve tweets for Donald Trump using hashtags like #DonaldTrump, #maga or the negative #NeverTrump we ended up with a lot more tweets than when we made a request for Hillary Clinton. So, we can conclude that people were talking about Donald Trump a lot more than they did about Hillary Clinton. It did not matter that a large percentage of these tweets were negative for Trump. It seems that there is no such thing like negative publicity.

From a technical point of view, we have to note that we were impressed with the variety of options that Twitter offers to download data. More specifically, the Search API is built in a way that developers can use easily and make the requests they want to get tweets. For analysts who want a different approach, Twitter offers the Streaming API which offers is real-time. In general, for analysts that want to gather data from social media we recommend that they do it with Twitter.

B. Future Work

This work demonstrates that social media analytics is a field with a bright future. There are a lot of different approaches that can be applied on similar experiments. In this section we discuss different approaches, solutions and tools that can be applied in social media analytics.

Firstly, sentiment analysis can be addressed using a different perspective. We used Naïve Bayes model in this work, but SVMs. Decision trees or other classification methods could have been used instead [27], [28]. There are cases that SVM might perform better than Naïve Bayes if we choose to base our analysis on other parameters. We can also try to address this problem using rule based approaches. Such approaches are not effective for unstructured data, but in our case the data are structured, so this method might be applied with success. We must further experiment to determine which the best way to approach the problem is.

In addition, the factor of importance for every tweet is not calculated. There are people that are considered influential and their tweets reach more people. The best approach would be to

assign weights for each tweet based on the retweets, likes and also the number of people following the person that posted it.

Also, there are some tools that can help to perform sentiment analysis more easily. One of the tools we could use is Lexicons. Lexicons are very helpful when performing sentiment analysis. They are useful resources of information about vocabulary. A very simple example of a Lexicon is a dictionary. In our case, we needed a lexicon that defines words as positive or negative. The intensity of a word is also important. There are words that express feelings more strongly than others. For example, the word "terrible" expresses a higher level of dissatisfaction than the word "bad". This is helpful regardless of what approach we have chosen to go with (Machine Learning or Rule Based).

It's not difficult to find well-structured lexicons. There are a lot of researchers and universities that design lexicons and make them available for everyone to download. Commonly used free available lexicons are the following:

- Inquirer
- MPQA
- LIWC
- SentiWordNet

The most commonly used Lexicon is SentiWordNet [29].

In conclusion, our method was proven to produce accurate results for predicting the US elections. There are many features and approaches that can improve our method and we plan to enhance our work in the future.

REFERENCES

[1] "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2017 (in millions)" Link: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ [Accessed 2/2/2018].

[2] S. Gosnell, "What are the most important states in 2016 presidential general election?" Link: https://www.quora.com/What-are-the-most-important-states-in-2016-presidential-general-election [Accessed 2/2/2018].

[3] S. Shepard, "The 11 states that will determine the 2016 election" 6/8/16, Link: http://www.politico.com/story/2016/06/donald-trump-hillary-clinton-battleground-states-224025 [Accessed 2/2/2018].

[4] K. Wegrzyn-Wolska and L. Bougueroua, "Tweets mining for French Presidential Election," Proc. 4th Int'l Conf. on Computational Aspects of Social Networks (CASoN), Sao Carlos, 2012, pp. 138-143.

[5] T. Mahmood, T. Iqbal, F. Amin, W. Lohanna and A. Mustafa, "Mining Twitter big data to predict 2013 Pakistan election winner," INMIC, Lahore, 2013, pp. 49-54.

[6] J. M. Soler, F. Cuartero and M. Roblizo, "Twitter as a Tool for Predicting Elections Results," Proc. IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 1194-1200.

[7] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation," Proc. Int'l Conf. on Control Communication & Computing India (ICCC), Trivandrum, 2015, pp. 638-641.

[8] A. Kumar, P. Dogra, and V. Dabas, "Emotion analysis of Twitter using opinion mining," Proc. 8th Int'l Conf. on Contemporary Computing (IC3) (IC3 '15). IEEE Computer Society, Washington, DC, USA, 2015, pp. 285-290.

[9] P. Ekman, "Basic Emotions," 1999 Handbook of Cognition and Emotion, Sussex, UK John Wiley & Sons, Ltd., 1999, pp.45-60.

[10] T. K. Das, D. P. Acharjya and M. R. Patra, "Opinion mining about a product by analyzing public tweets in Twitter," Proc. Int'l Conf. on Computer Communication and Informatics, Coimbatore, 2014, pp. 1-4.

[11] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Proc. 4th Int'l AAAI Conf. on Weblogs and Social Media, pp. 178-185, 2010.

[12] A. Jungherr, H. Schoen, O. Posegga and P. Jurgens, «Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support», Social Science Computer Review 35(3) February 2012, pp. 1-21.

[13] Twitter for developers https://dev.twitter.com/index [Accessed 2/2/2018].

[14] P. Norvig and S. Russell, Artificial Intelligence: A Modern Approach (3rd ed.). Prentice Hall, 2009.

[15] S. Ray, "6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)": https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [Accessed 2/2/2018].

[16] Twitter Sentiment Analysis Training Corpus (Dataset), Link: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/ [Accessed 2/2/2018].

[17] "TextBlob: Simplified Text Processing", https://textblob.readthedocs.io/en/dev/ [Accessed 2/2/2018].

[18] The Huffington Post, "Our @pollsterpolls model gives @HillaryClinton a 98.1% chance of winning the presidency", 7/11/16, Link: https://twitter.com/HuffPost/status/795663593689808896 [Accessed 2/2/2018].

[19] NY Times "Latest Election Polls 2016", 8/11/2016 Link: https://www.nytimes.com/interactive/2016/us/elections/polls.html [Accessed 2/2/2018]

[20] L. Graham, "85% chance of Clinton winning the US election, say UK betting firms", 26/10/2016, Link: http://www.cnbc.com/2016/10/26/85-chance-of-clinton-winning-the-us-election-say-uk-betting-firms.html [Accessed 2/2/2018].

[21] New York Times, "Our presidential forecast, updated", 20/10/2016, Link: https://twitter.com/nytimes/status/789083772205600768?lang=el [Accessed 2/2/2018].

[22] C. Tjortjis, "On US elections", 7/11/2016, Link: https://www.facebook.com/Christos.tjortjis/posts/10155428092860898, 9/11/16, Link: https://www.linkedin.com/pulse/us-elections-christos-tjortjis [Accessed 2/2/2018].

[23] N. Scola, "How Facebook, Google and Twitter 'embeds' helped Trump in 2016", 26/10/2017, Link: https://www.politico.com/story/2017/10/26/facebook-google-twitter-trump-244191 [Accessed 2/2/2018].

[24] P. Burnap, R. Gibson, L Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using Twitter to predict the UK 2015 General Election", Electoral Studies, 41, 2016, pp. 230-233.

[25] V. Isotalo, P. Saari, M. Paasivaara, A. Steineker, and P.A. Gloor, "Predicting 2016 US Presidential Election Polls with Online and Media Variables", Designing Networks for Innovation and Improvisation, . Springer International Publishing, 2016, pp. 45-53.

[26] M.P. Cameron, P. Barrett, and B. Stewardson, "Can social media predict election results? Evidence from New Zealand", Journal of Political Marketing, 15(4), 2016, pp. 416-432.

[27] P. Tzirakis P. and C. Tjortjis, "T3C: Improving a Decision Tree Classification Algorithm's Interval Splits on Continuous Attributes", Advances in Data Analysis and Classification, Springer, Vol. 11, No. 2, 2017, pp. 353-370.

[28] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, and J. Keane, "Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction", Information Systems Frontiers Journal, Springer, Vol. 11, No. 4, 2009, pp. 449-460.

[29] SentiWordNet, Link: http://sentiwordnet.isti.cnr.it/ [Accessed 2/2/2018].

.