

Opinion Polls: Election Result Prediction Using Social Networking Sites: A Review

Falak Jan^{1,*}, Afaq Alam Khan²

¹Student, Department of Information Technology, Central University of Kashmir, Ganderbal, Jammu and Kashmir, India

²Department of Information Technology, Central University of Kashmir, Ganderbal, Jammu and Kashmir, India

Abstract

The impact of social media is increasing and is being embedded in daily life as it provides a platform for people to express their opinions, sentiments and feelings towards any particular topic. So the data from social media is used by researchers as a data source for opinion mining or sentiment analysis. It can be used to discover political preference of users regarding elections as it has become the most important tool for communication over the internet and provides with the public reactions. In this study, we discuss the most relevant and recent research, the approaches and techniques that are followed for predicting election outcome. This study also presents the results that were recorded in the different experiments by various researchers.

Keywords: Social media, opinion mining, sentiment analysis, election prediction

***Author for Correspondence** E-mail: falakmaraxi@gmail.com

INTRODUCTION

The field of opinion mining or sentiment analysis is developing due to its wider practical applications. This field analyzes the text to find the underlying drivers that are behind the public sentiments that can be emotions, sentiments, assumptions towards different issues, subjects, fields, individuals etc. These drivers can be used to unfold the acute areas of strengths and weaknesses. In order to extract knowledge from bulk of data, data mining has become a necessity. Today, most of the organizations use this field for making plans. Owing to the restrictions that summarizing information faces, and the detection of opinion from among these human analyses, suffers considerable amount of bias. Thus there is a need of automated opinion mining to overcome this biasness.

The second related area is forecasting or prediction. People try to predict different economic and political events based on previous and current data and mostly by analyzing the trends; and the typical example is, predicting election outcome. This area is

concerned with future events as it shows the probability of coming events.

For years, different methods like polls have been used by researchers for measuring the intentions and opinions of individuals. Due to the limitations involved in traditional polls including the human effort involved, in addition to being tedious, costly and unable to predict the election outcome accurately, the automated systems are thus needed. Trying to resolve the certainty and the high cost issues, data from social media can be used to predict the election outcome. Social media has become the most important source for opinion mining as it provides adequate data source for analyzing important topics and its related public reactions. In addition, it enables the users to express their perception and sentiments as much as possible. This task is challenging as it has to ensure that the sample has representative distribution.

In this study, we are trying to concise various related state of art approaches used in different but related studies together into a single point.

We discuss the state of art techniques and approaches along with their advantages and disadvantages.

RELATED WORK

The impact of social media like Twitter, Facebook etc. has grown and become the most prominent tool for communication and is being used as alternative form of political debate and gathering information. Different studies focus on the impact of social networking sites on the population for different issues such as politics in order to get their political view. Twitter allows the researchers to collect the tweets using search API that is available to public.

One of the applications of using the data from social networking sites is to predict election results. Standard methodologies are used by researchers to predict the outcome of an election. Sentiments from tweets are used to infer the political preferences to figure out voting preferences of different people. In this section, we review the related work regarding election prediction from social networking sites using opinion mining.

Methods Used to Predict Voting Preferences

Different methods are used by pollsters to predict voting preferences:

One approach is using traditional opinion polling consisting of telephone poll based on small data size (from 1000-1500) and online polls that take larger data size (from 2000-5000). But these methods are found to be demographically biased.

Celli *et al.* used forecasting based on sentiments analysis and collected data using some hand crafted rules such as hashtags, keywords and mentions and then mapped them to their relevant labels [1]. Mapping of posts using these rules was found not to be enough to infer the opinion of author because sometimes an author may use a particular hashtag/keyword/mention only to criticize.

So they described two systems to predict author's opinion:

1. Agreement/disagreement classification; and
2. Sentiment polarity.

Agreement/Disagreement Classification

Agreement/disagreement classification, based on stylometric features that uses statistical techniques to analyze literary styles such as Lexical, Syntactic or Word based features. Lexical based features are collected in terms of characters, e.g., total number of capital letters, total number of white spaces, total number of alphabetic characters etc. While word based features include word based distribution, e.g., average number of words per sentence, indicating the preferences of particular word; and syntactic features include POS (Part Of Speech), capturing author's style of writing at the sentence level that is obtained from people's different ways of organizing each sentence. This model considered the ratio of characters and word ngrams (sequence of tokens). They assigned the features with the values between 0 and 1.

Sentiment Polarity

While the prediction using sentiment polarity was lexicon based. Each post was labeled either positive, negative or neutral. Positive lexicons expressed the desired state while the negative lexicon expressed undesired states. All these lexicons are collectively called as opinion lexicons which play an important role in opinion mining. The system got the accuracy of 68.7% on Movie Reviews 2.0 data set.

So, it was found that NLP techniques like opinion mining are beneficial for opinion polling and the referendum events can be forecasted with high accuracy. However the accuracy of prediction depends on many factors like the time period as well as the method used for analysis.

They compared NLP techniques with traditional polling and found that the NLP techniques produce the outcome with high accuracy as they analyze larger sample size of a minimum of 80000 to 1 lakh posts per day; while traditional polls can analyze only a smaller sample size on average 1000 to 4000 individuals.

NLP techniques try to deduce the opinions of individuals while as traditional polling enquires people about their opinions. Also the data available on social networking sites cover larger and diverse range of geographic locations.

Salunkhe *et al.* used different techniques to predict the outcome of election by deducing sentiment of data collected from twitter about the candidate [2]. They collected data from twitter using either of the two methods, by selecting tweets based on the labels or through streaming API. All these tweets were then inserted into database. Then they performed preprocessing in order to increase the quality of text which involved various steps like lower case conversion, removing punctuation and numbers, steaming and stripping white spaces etc. They used two approaches for sentiment analysis:

- Supervised machine learning approach which has finite set of classes for the classification. Some of the common classification algorithms are Naïve Bayes, Support Vector Machine etc. But they classified tweets using Naïve Bayes classifier into polarity and emotions.
- And the second approach used for sentiment analysis was Lexicon based in which they opted for dictionary based approach for classification using eleven variables: sadness, anxiety, positive words, negative words, anger, positive hashtags, negative hashtags, certainty, work, achievements. The collected words were then classified according to these variables to calculate the sentiment score that helped to predict the outcome of election.

Tunggawan *et al.* described a simple data preprocessing technique in order to label and train the predictive model for 2016 US Presidential Election using data from twitter [3]. They collected 371,264 tweets since December 16, 2015 until February 29, 2016 using Streaming API with the search keyword as #Election2016. This data was saved as JSON files in which each line represented a tweet with 26 main attributes, such as ID, created_at, text, retweet_cnt and language.

They used only the content of two attributes created_at and text as their main focus was the sentiment towards the candidate at a particular time, excluding the location of candidate and other information.

After data collection, preprocessing was performed by removing URLs, pictures except Hashtags, Mentions and retweets and by filtering tweets that included candidate's name. Then data was labelled by 11 English annotators and the label validity was determined by means of majority rule. The sentiments from data were predicted using Bayesian model which included Model Training, Model Accuracy Test and Prediction Accuracy Test. Data collected since December 16, 2015 until February 2, 2016 was used to train the model using Naïve Bayes Classifier. 10-fold cross validation was used to test the accuracy of model that was calculated by checking confusion matrix and F1 score. The model used tweets collected from February 3 until 9, 2016 as the prediction model in which the prediction was made using two steps:

- First by calculating positive sentiments from tweets, and
- then ranking the candidates by the number of sentiments that are positive.

They calculated the error rate by comparing their calculated ranks with the results on RealClearPolitics.com. Their model achieved 95.8% accuracy for model accuracy test and 54.8% accuracy for prediction accuracy test.

Choy *et al.* there was an intense battle between the two candidates in US Presidential Election 2012. So they presented a model that incorporated both socio-demographic as well as census information in addition to sentiment analysis in order to predict the election outcome [4]. They collected 7,541,470 tweets for 3 months starting from August 12, 2012 until October 31, 2012. Then a generic purpose corpus was used to automatically draw out the sentiments from data. The collected data included census information as well as demographic data to reduce the biasness in the online data. They also used the combination of age, computer literacy in addition to the prior election information to

determine the vote share. Two additional pieces of information were used: percentage of US twitter users, and state by state breakdown of internet usage which provided with two models:

- The first model assumed that the twitter information reflects the sentiments of overall population. The absolute error for this model was calculated to be 0.04%. However the model performed less for state level model.
- Whereas the second one modeled on the basis of previous political affiliation with absolute error of 1.65% when compared with baseline, which was undesirable.

Rill *et al.* proposed a system which could identify the upcoming topics faster than other channels called as PoliTwI for 2013 parliamentary elections in Germany [5]. Hashtags are usually used to mark the new topics as Top Topics, in addition to it they added a sign to each hashtag to represent the polarity called as *sentiment hashtag*. Their basic idea was to compute the temporal change in the number of tweets by comparing the current number of tweets with hashtag to the previous number of tweets taking into consideration the standard deviation. They assigned values to the topics between -1 and 1 to decide whether the particular topic is a Top Topic and only those topics were selected whose value was greater than zero. For understanding the unstructured text they used a semantic approach and thereby building a relation graph for upcoming political topics using sentiment hashtags. Their basic idea was to show how relation between topic and its polarity could increase the knowledge for concept level sentiment analysis.

Berminham *et al.* combined the two approaches, supervised learning and volume based measure for analyzing sentiments in order to predict election outcome using Irish General Election, 2011 as a case study for investigation [6]. They used Twitter Tracker to collect the data from twitter that featured various abstractive and extractive approaches for summarization. They presented two approaches for examining the sentiments in a particular topic, inter-party and intra-party. It

was found that volume based measure strongly indicated the sentiments followed by inter-party sentiments.

Khan *et al.* proposed a hybrid approach for classification of tweets in real time to increase the classification accuracy [7]. The different techniques that they employed in this approach are Data acquisition, Pre-Processing, Classification, and Evaluation. They used three classifiers for tweet classification: Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC) and SentiwordNet classifier (SWNC). They used EEC for classification of emoticons, IPC for classification of the word, and SWNS for classification of Tweets using SentiwordNet Dictionary. They used six Twitter datasets for experimentation and produced an accuracy of 85.7% with 82.2% recall and 85.3% precision.

Ortigosa *et al.* presented a hybrid approach of lexical-based and machine-learning techniques using SentBuk for Classification of Facebook Status posts [8]. They used J48, Naïve-Bayes and Support Vector Machines to build the classifier. The performance of this classifier was excellent with an accuracy of 83.27%. This approach can be used in e-learning systems and recommended systems.

The generalized system implementation for predicting election outcome can be shown as in Figure 1.

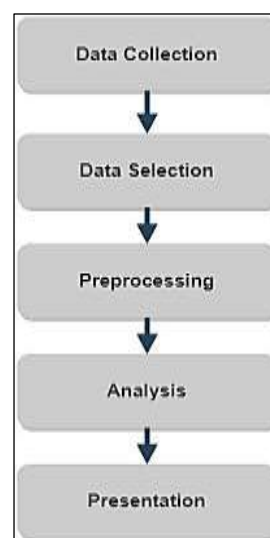


Fig. 1. Overview of Election Prediction.

Summarization of the Previous Work Done: Results are shown in Table 1.

Table 1: Summerization of Results Obtained.

Prior Work	Year	Dataset	Approaches Used	Performance (Accuracy %)
Celli <i>et al.</i> [1]	2016	Italian CorEA corpus	NLP Techniques	68.7
Salunkhe <i>et al.</i> [2]	2017	Twitter Data	Machine Learning Approach (Naïve Bayes) and Lexion Based Approach (dictionary based approach)	Not Available
Tunggawan <i>et al.</i> [3]	2016	Twitter Data	Naïve Bayes Classifier and 10-fold cross validation	95.8
Choy <i>et al.</i> [4]	2012	Generic purpose Corpus	Sentiment analysis	Absolute error for Model 1= 0.04 Absolute error for Model 2=1.65
Rill <i>et al.</i> [5]	2014	Google Trends	Concept-level sentiment analysis	Correlation result for <i>Neuland-new territory</i> between PoliTwii and Google Trends ➤ PoliTwii: corr ₀ : 0.68 R ² ₀ : 0.47 p-Value ₀ : 3.1e-0.5 ➤ Google Trends: corr ₁ : 0.81 R ² ₁ : 0.65 p-Value ₁ : 6.7e-0.8
Bermingham <i>et al.</i> [6]	2011	Twitter Data	Sentiment analysis using supervised learning, Volume based measure and Sentiment TF-IDF	Mean Absolute Error: 3.76
Khan <i>et al.</i> [7]	2014	6 Twitter Datasets	Data acquisition, Pre-Processing, Classification, and Evaluation	85.7
Ortigosa <i>et al.</i> [8]	2014	Facebook Status posts	Lexical-based approaches and machine-learning techniques (Naïve-Bayes and Support Vector Machines)	83.27

APPROACHES

The different approaches used in sentiment analysis for election prediction are (Figure 2):

1. Supervised Learning.
2. Unsupervised Learning.
3. Opinion Lexicon Expansion.
4. Aspect based sentiment analysis.

Supervised Learning

Supervised learning approach uses finite set of classes for classification. The model learns how to create a mapping between the input and the particular output based on what is learns from input labeled data set. Some of the algorithms used for classification are:

- i. Support Vector Machine: Support Vector Machine is one of the learning models that analyze data for classification in accordance with the predefined labels.
- ii. Naïve Bayesian Classification: This is another model used for classification and is based on Bayes probability theorem.

Unsupervised Learning

Unsupervised learning approach does not need to build set of training data in order to classify; instead it uses the hidden features and map the data into different classes. Thus the model is not supervised; it works on its own to discover hidden information that may not be visible to the human eye. It uses machine learning algorithms to draw conclusions on unlabeled data.

Opinion Lexicon Expansion

Opinion words or sentiment words can be used in various tasks for sentiment classification which can be either positive or negative. Positive opinion words express the desired state while the negative opinion words express the undesired state.

- i. Dictionary based approach: In this approach first the opinion words are collected manually to create a set and then their synonyms and antonyms are searched

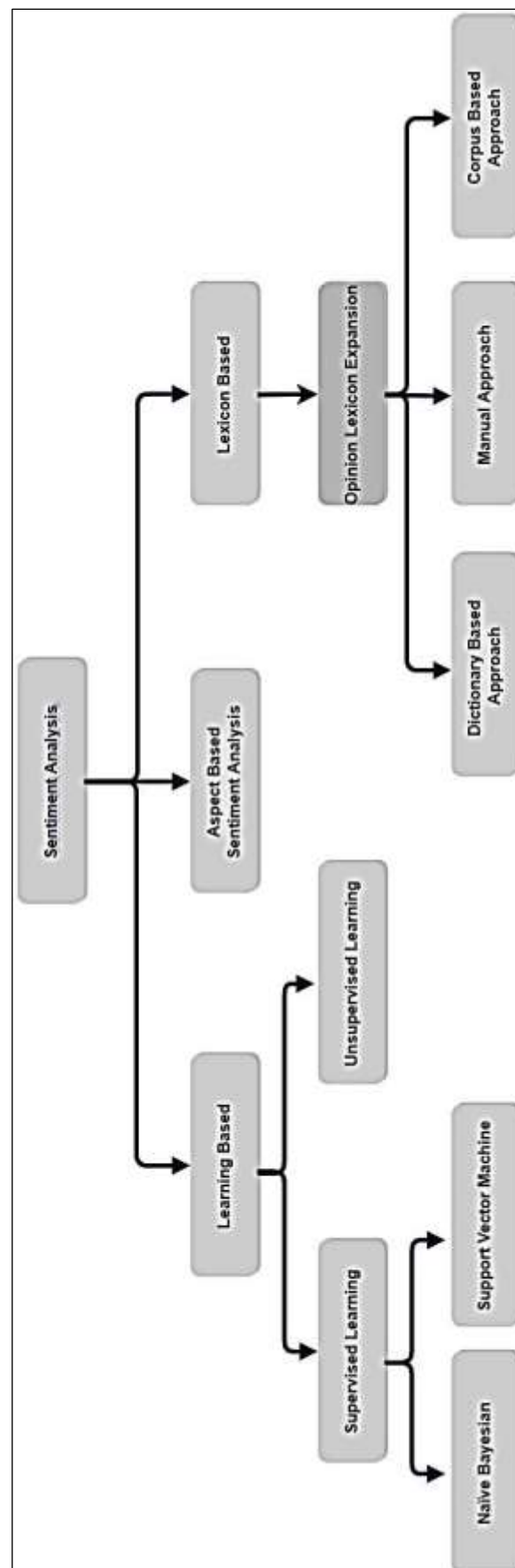


Fig. 2. Approaches to Sentiment Analysis.

- ii. using online dictionary to enrich the seed list. This process of adding words stops when no matching word is found.
- iii. Manual approach: This approach is inefficient and the most time consuming one but can be used to infer the final judgements.
- iv. Corpus based approach: This approach relies mostly on the syntactic patterns and the list of opinion words to infer the polarity of words but is difficult to prepare. The focus is to find the domain, opinion words that are specific to a particular context and their orientation with the help of the domain corpus.

Aspect Based Sentiment Analysis

This approach analyzes different attributes present in the posts and classifies them as positive, negative or neutral according to the particular aspect in the sentence. First, the aspects that are to be evaluated are extracted after that the polarity on different aspects, is determined using classification.

CONCLUSION

In this study, we discussed how data from social media can be used to predict the outcome of election. This field is emerging because people express their opinions and sentiments regarding everything, from the details of everyday life to the opinion about the issues like their political preferences. So, we present a survey paper to explore this field. In this study we have also discussed the relevant state of art techniques and approaches along with their merits and demerits. Each approach and technique in different domain has been explained elaborately. We have also summarized the results obtained and the data set that was used by different researchers as reported in Table 1.

REFERENCES

1. Celli F, Stepanov E, Poesio M, *et al.* Predicting Brexit: Classifying Agreement is Better than Sentiment and Pollsters. In *Proceedings of the Workshop On*

- Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*. 2016; 110-118p.
2. Salunkhe P, Deshmukh S. Twitter Based Election Prediction and Analysis. *IRJET*. Oct 2017; 4(10): 539-544p.
3. Tunggawan E, Soelistio YE. And the Winner is...: Bayesian Twitter-Based Prediction on 2016 US Presidential Election. In *IEEE Computer, Control, Informatics and its Applications (IC3INA), 2016 International Conference on*. Oct 3, 2016; 33-37p.
4. Choy M, Cheong M, Laik MN, *et al.* US Presidential Election 2012 Prediction Using Census Corrected Twitter Model. *arXiv preprint arXiv:1211.0938*. Nov 5, 2012.
5. Rill S, Reinel D, Scheidt J, *et al.* Politwi: Early Detection of Emerging Political Topics on Twitter and the Impact on Concept-Level Sentiment Analysis. *Knowl Based Syst*. Oct 1, 2014; 69: 24-33p.
6. Bermingham A, Smeaton A. On Using Twitter to Monitor Political Sentiment and Predict Election Results. In *Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011)*. 2011; 2-10p.
7. Khan FH, Bashir S, Qamar U. TOM: Twitter Opinion Mining Framework Using Hybrid Classification Scheme. *Decis Support Syst*. 2014; 57: 245-257p.
8. Ortigosa A, Martín JM, Carro RM. Sentiment Analysis on Facebook and Its Application to E-Learning. *Comput Hum Behav*. 2014; 31(1): 527-541p.

Cite this Article

Falak Jan, Afaq Alam Khan. Opinion Polls: Election Result Prediction Using Social Networking Sites: A Review. *Journal of Artificial Intelligence Research & Advances*. 2019; 6(2): 1-7p.