IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTING ELECTIONS THROUGH SENTIMENT ANALYSIS ON TWITTER

[1]S THAIYALNAYAKI , [2]G GOWTHAMI, [3]G SRIHARI, [4]G NARASIMHA RAO , AND [5]J MIDHUN

[1]Associate Professor, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073.

[2,3,4,5]Students, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073.

ABSTRACT— Social media's recent explosion has given individuals a strong platform on which to express their ideas. To understand user orientation and make better judgments, businesses (or similar entities) must determine the polarity of these viewpoints. One such use is in politics, where political organizations must comprehend public opinion in order to choose their approach to campaigning. Many people believe that sentiment analysis of social media data is a useful technique for keeping track of user inclinations and preferences. In order to perform sentiment analysis, supervised learning techniques—which are utilized in well-known text classification algorithms like Naive Bayes and SVM—need a training data set. Both the volume and the quality (features and contextual significance) of the labeled training data affect the algorithms' accuracy. Most applications use cross domain sentiment analysis, which loses out on features pertinent to the target data, because they lack sufficient training data. Consequently, this reduces the overall text classification accuracy.

Keywords: Vader, Twitter, sentiment analysis, text classification, training data, labeling, Passive aggression Classifier

I. INTRODUCTION

Social media's recent explosion has given individuals a strong platform on which to express their ideas. There is active usage of social media sites like Facebook, Twitter, and Google+ to exchange ratings, reviews, and suggestions.

The authors of offer suggestions for how social studies and marketing might make active use of this enormous amount of data. Utilizing the wide range of data accessible on the aforementioned platforms, political campaigns have been able to get insights about user opinions and subsequently craft their marketing strategies. Politicians' expenditures on social media campaigns just before elections, coupled with discussions and debates between opponents and supporters, serve to strengthen the premise that user-posted views and opinions influence election outcomes.

The author's sentiment toward a political party or an election candidate can be ascertained using a variety of sentiment analysis methods. "Tweets" are 140-character messages that users can send and read on Twitter, an online social networking platform. The idea of hash tags makes this platform more intriguing.

In addition to the brief messages, users can classify their Tweets and make them appear more readily in Twitter Search by adding the hash tag sign "#" before a pertinent term or phrase. Since a hash tag can express an opinion or an emotion, using them helps to simplify the text classification problem. The official hashtag for Republican Presidential Candidate Donald Trump, for example, is #MakeAmericaGreatAgain. Support for this candidate would be indicated by any tweets that contain this hash tag. Additionally, it has been observed that as technology has advanced, internet platforms have become more dependable and affordable for outcome prediction. It has recently come to light that traditional polls can fall short of providing a clear and accurate forecast. Therefore, in an effort to increase the accuracy of election outcome predictions, scientists and academics have focused their attention on looking through and evaluating web data, such as blogs or user activity on social networks. In addition, traditional survey methods are prohibitively costly, but data on the internet is readily available and may be obtained for free. According to, With over a million messages sent every hour, the public timeline that houses the tweets of every social media user on the planet is a massive real-time information flow. The main objective of microblogging was to offer updates on one's personal situation.

II. RELATED WORK

The many text categorization strategies utilized in for situations without training data are listed in this section. This covers sentiment analysis across domains and unsupervised learning. The methods by Taboada et al, Harb et al, and Turney are discussed for unsupervised learning. Turney computes the semantic orientation (point-wise mutual information with respect to a positive and a negative seed word) of verbs and adjectives in a sentence. By summing the independent values for semantic orientation, he ascertains the overall polarity. They used this method to attain a 74% accuracy rate. Harb et al. established relationships for positive and negative phrases using the Google search engine.

The methods of Wu and Tan and Liu and Zhao for cross-domain sentiment analysis are examined. Wu and Tan employ the following two-stage framework: Using a graph ranking algorithm, an association is initially established between the source and the target domain. Next, a few of the target domain's best seeds were chosen. In the second step, each document's sentiment score was determined using its essential structure, and the target-domain documents were subsequently labeled in accordance with these scores.

Additionally, Liu and Zhao suggest a two-stage approach. They translated a feature from the source domain to the target domain in the first step of their approach by using a feature translator.

Using the data from the source domain to train a classifier, they classified the unlabeled data in the target domain in the second stage.

Compared to supervised learning approaches, the total accuracy of the two previously mentioned methodologies has been less than 70%. As such, it supports the idea that, as demonstrated in obtaining highly accurate text classification results requires a training data set that is both accurate and contextually relevant. But the authors of only manage to obtain a sparse data set of 1000 tweets, which is insufficient to meet the quantitative need of a supervised learning algorithm.

III. DATA SET CREATION

A. Data Gathering

Twitter information for two candidates, Donald Trump and Joe Biden was gathered on March 12 and 15, 2024.

We retrieved pertinent information on the presidential contenders using the Twitter Streaming API. Developers can access Twitter's global stream of Tweet data with minimal latency thanks to the Streaming APIs. The names of the presidential candidates as well as additional keywords like "Democrats" and "Republicans" were entered as input parameters for the streaming routines. JSON-formatted tweets that matched the specified parameters were returned. In essence, the JSON result was made up of key-value pairs. A few keys were made at, among other places, id, screen name, location, and retweeted. Only the tweet's body was extracted from the JSON answers, which were then saved in a CSV file.

TABLE I

ESTIMATES THE DATA FOR CANDIDATES FOR PRESIDENT

Candidate	*Total Tweets*
Donald Trump	60, 473
Joe Biden	61, 121

B. Preprocessing the Data

Special characters like '@' and URLs were now eliminated from the tweets in order to reduce noise. In order to improve the classifier accuracy, we also employ the TF- IDF (term frequency - inverse document frequency) technique in the Machine Learning modules to identify phrases that are more closely associated with moods.

C. Labeling Data

Our two-step methodology for producing a labeled training data set is provided in this section. This two-stage architecture facilitates the creation of a data set that satisfies the needs of a supervised learning algorithm by being contextual and non- precise at the same time.

Step 1: Using hash tag clustering for manual labeling

The human labeling of the Twitter data is the initial step in this approach. But manual labeling of the whole Twitter data collection is not necessary. We present a method we call grouping of hash tags. When mining Twitter for data, users frequently come across several tweets with the same hash tag. Take the official hashtag #MakeAmericaGreatAgain, for example, which represents US Presidential Candidate Donald Trump. Now that this is Donald Trump's official hash tag, it goes without saying that everybody who tweets with it supports Trump. As a result, any tweet that uses the hashtag #MakeAmericaGreatAgain needs to be associated with Trump in a positive light. Thus, thousands of tweets with the same hash tag can be automatically classified via a code just by linking a label to the hash tag.

It is required to sort the hash tags in decreasing order of frequency prior to employing this strategy. By doing this, it will be ensured that hash tags with higher frequencies are labeled before those with lower frequencies. Developers or analysts may even decide not to label hash tags with lower frequency or those that are unclear, such as #MakeAmericaGreatAgain, depending on the application, since they will be covered in our next section.

The possibility that a candidate is connected to a scheme or initiative increases the importance of manual labeling. Using a cross-domain data collection, it is not possible to identify tweets containing the hashtag #joebiden negatively for Joe Biden in the joe Biden scandal.

Step 2: Adding the VADER hashtag to the remaining tweets

A vocabulary and rule-based sentiment analysis tool that is especially well-suited to the sentiments conveyed on social media is called Vader (Valence Aware Dictionary and Sentiment Reasoner). It was essentially designed as a sensation intensity polarizer by Hutto and Gilbert. Vader takes a sentence as input and returns a percentage value for each of the three categories: positive, neutral, negative, and compound, which represents the overall polarity of the text.

TABLE II

VADER SENTEMINT ANALYSIS EXAMPLES

Sentence	*comp*	*pos*	*neg*	*neu*
He is smart and funny	0.83	0.75	0.0	0.254
A horrible book	-0.82	0.0	0.791	0.20
It sux, but I will be fine	0.22	0.274	0.195	0.53

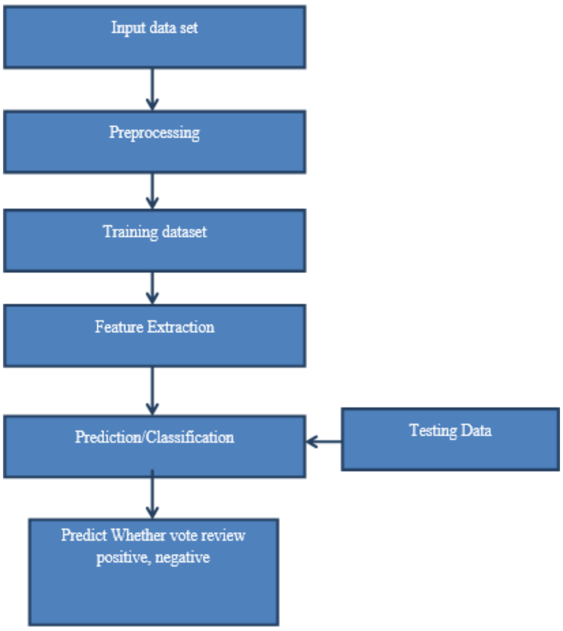

Fig. 1. Data flow diagram

Three instances of sentences examined using Vader are shown in the above table. First sentence is quite positive, second statement is very negative, and third sentence is neutral. Sentences in a training data set should be clearly classified as positive or negative in order to facilitate sentiment analysis. Therefore, the training data set should only contain statements with a compound value of >=0.8 (very positive) or <=-0.8 (extremely negative), depending on our observations. The rest sentences can be deleted. Vader's Python implementation is easily accessible as an open source project on GitHub.

As a result, a training data set for Twitter can be produced using the two-stage framework previously suggest, For example, the user's demands can determine whether to raise or lower the 0.8 threshold in stage 2. Similarly, stage 1 can be immediately removed and Vader can be used to label every sentence in situations when the frequency of sentences for each hash tag is extremely low (such as 10 or 20 sentences).

TABLE III

LABELED DATASET USING 2 STAGE FRAME WORK

Candidate	*Stage I*	*Stage II*	*Total*
Donald Trump	17166	7321	24487
Joe Biden	17115	14435	31550

The original data set comprised approximately 60,000 tweets pertaining to specific politicians. For the training data set, we receive about 30,000 tweets after classifying the data set using the two-stage framework. This is as a result of the two-stage framework's extremely high (80%) Vader threshold. Consequently, unclear or extremely neutral tweets are removed, improving the caliber of the training data set.

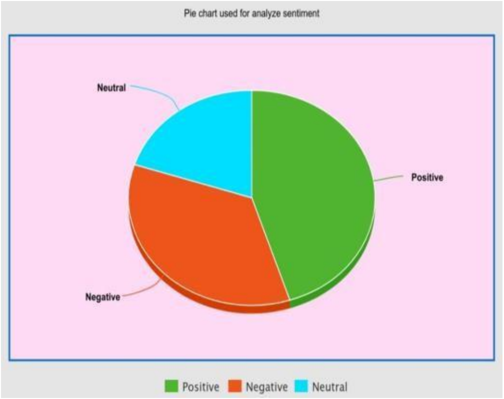

Fig. 2. Pie chart used to analyze sentiment

D. Algorithms

There are numerous methods available now for sentiment analysis in particular and natural language processing in general. To ascertain the polarity of tweets, we employed two algorithms: Support Vector machines and Multinomial Naive Bayes. Scikit Learn and NLTk are two packages available in Python that can be used to implement the aforementioned techniques. As mentioned in 2.3.1, we tested both of the packages for sentiment analysis on the hand labeled data set. The following table shows the accuracy for both algorithms:

TABLE IV
ACCURACY FOR ALGORITHMS OF SENTIMENT ANALYSIS

Package	*Algorithm*	*Accuracy*
nltk	MNB	0.54
nltk	SVM	0.58
Scikit-learn	MNB	0.97
Scikit-learn	SVM	0.99

The Scikit-learn package's SVM method has the best classification accuracy, as can be seen in the above table. For our final model, we therefore employ the SVM algorithm—specifically, the one from Scikit- learn.

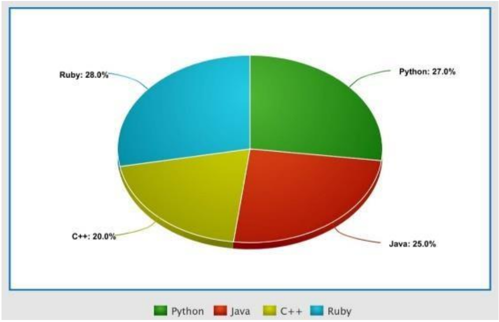

Fig 3. Pie chart was utilized in programming tools.

IV. RECOMMENDED MODEL

Now that we have a two-stage labeled dataset, we can utilize it to train a supervised machine learning model that can analyze public mood and forecast election results. To create the training and testing sets, we divided the dataset into 80:20 ratios.

TABLE V

TRAINING AND TESTING DATA FOR CANDIDATES

Candidate	*Training*	*Testing*	*Total*
Donald Trump	19589	4898	24487
Joe Biden	25240	6310	31550

A. Design

We carry out multistage classification for our suggested model to determine whether a tweet is positive or bad regarding one of the candidates running for office. To begin with, we categorize the tweet according to the candidate it pertains to or addresses. An 'entity classifier' is the initial classifier; it divides a generic stream of data into the appropriate entities. In the subsequent phase, the categorization is carried out according to the tone of the text concerning that specific applicant. As a result, each candidate is linked to a classifier. The complete data set labeled by the entities is used to train the entity classifier. Data sets specific to the sentiment classifier's candidate are used to train it.

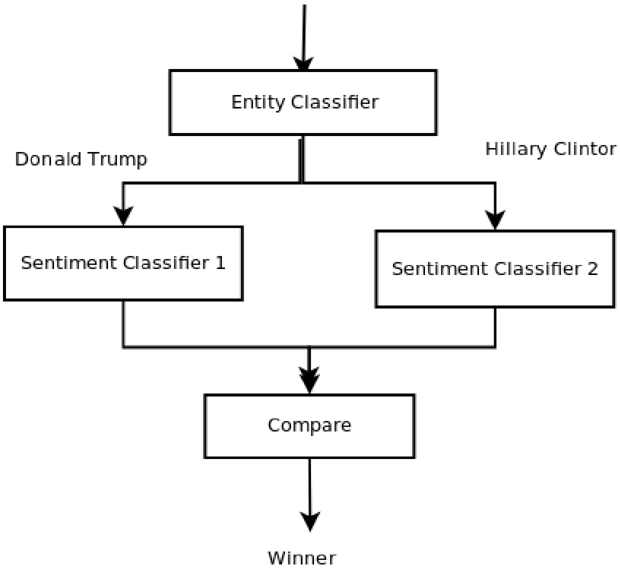

Fig. 4: Election Result Prediction Model

B. Execution

We compared the following classifiers' performances on our preprocessed, labeled data set in order to build the supervised classification model design. The following are some common uses for these classifiers in text-based classification:

TABLE VI
SUPERVISED CLASSIFICATION TECHNIQUES COMPARISON

Classification Technique	*F-1 Score*
SVM Linear Kernel	0.97
SVM – rbf kernel	0.39
SVM – liblinear	0.97
Naïve Bayes – MultinomialNB	0.94

We chose the SVM with linear kernel as our entity and sentiment classifier based on the F-1 score metric.

- After categorizing 'Joe Biden' and 'Donald Trump' using training data of 50,433 tweets and testing data of 5,603 tweets, the entity classifier produced an accuracy of 0.98.
- Using testing data of 4,898 tweets of "Donald Trump" and training data of 19,589 tweets, the sentiment classifier produced an accuracy of 0.99.
- When we used training data consisting of 25,240 tweets and testing data consisting of 6,310 tweets mentioning "Joe Biden," the sentiment classifier produced an accuracy of 0.97.
- Table VI shows the results of the testing data for both sentiment classifiers.

C. Consolidation

The winner was determined by calculating the Positive versus Total count ratio (PvT Ratio), which was determined as

|P| / |T} = Ratio (1)

Here, P represents the tweets that the candidate's sentiment analyzer determined to be favorable, and T represents all the tweets that the entity classifier determined to be linked to the candidate.

TABLE VII
PvT RATIO FOR CANADIDATES

Candidate	*Positive*	*Negative*	*Total*	*PvT Ratio*
Donald Trump	1378	2410	4851	0.365
Joe Biden	2681	2170	4851	0.554

Since the data set count may be skewed towards one contender over another, the direct count of favorable tweets cannot be used as a criteria to choose the winner. In the event that 30,000 tweets are mined for Joe Biden, of which 9,000 are positive, and 50,000 tweets are mined for Donald Trump, of which only 10,000 are positive, the results of a direct comparison of the positive tweets would be inaccurate because the percentage of positive tweets for Clinton is significantly higher.

Therefore, calculating the PvT Ratio—the percentage of positive tweets for each politician—will provide a reasonable indication of each candidate's level of popularity.

V. CONCLUSION

Predicting election results with social media presents a number of challenges. To tackle the deficiency of training data for text classification, we initially introduce a two-phase framework in this study. Lastly, we present our model for predicting election results, which makes use of the labeled data produced with our system. Even though our model by itself would not be able to forecast the outcomes, when paired with other statistical models and offline methods (such as exit polls), it becomes an essential component.

We used the suggested model to a dataset that was produced over the course of three days of Twitter mining.

It is possible to expand this model in the future and develop an automated framework that can mine data for months, as predicting election results is a continuous process that necessitates analysis over longer periods of time. Newly mined data should have its features extracted and compared to the current feature set. The new and old features can be compared using a similarity metric. The two step framework should only be used to label the newly mined data when the metric value exceeds a threshold. Therefore, we advise developing an active learning model in which the model suggests the labels for the data. This would ensure that contextual relevance is maintained without sacrificing labeling effort.

VI. REFERENCES

1. Alexander Pak and Patrick Paroubek." Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of the Seventh International Conference on Language coffers and Evaluation(LREC ' 10), may 2010.

2. Yang andF. Zhou," Microblog Sentiment Analysis Algorithm Research and perpetration Grounded on Bracket", 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science(DCABES), 2015.

3. IJRCCT, vol. 1, no. 4, pp. 133–138, 2012; M. Hajmohammadi, "Lack of Training Data in Sentiment Classification: Current Solutions“

4. P. D. Turney, "Applying semantic orientation to unsupervised review classification: thumbs up or thumbs down?" was presented at the 40th Annual Meeting of the Association for Computational Linguistics in Philadelphia, Pennsylvania in 2002.

5. "Lexicon-based methods for sentiment analysis," by M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, published in Comput. Linguist., vol. 37, pp. 267-307, 2011.

6. "A two-stage framework for cross-domain sentiment classification," Q. Wu and S. B. Tan, Expert Systems with Applications, vol. 38, pp. 14269-14275, Oct. 2011.

7. Dev.twitter.com, "The Streaming APIs | Twitter Developers", 2016.

8. Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013 IEEE, Neethu, M. S., and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques."