Opinion mining and Sentiment analysis of TELANGANA election on twitter data

Kishan Kumar Sahu[1], Vanitha kakollu[2]
[1,2]*GITAM (Deemed to be University), Visakhapatnam*

Abstract

The opinion mining and the sentiment analysis of the network comment are the key points of the text analysis. By excavating the comment information of public opinion on Election, the main focus is to predict Election results. Election is conducted to view the public opinion, where group of people choose the candidate by using votes; many methods are used to predict result. Many agencies and media companies conduct pre poll survey and expert views to predict result of election. In this, twitter is used to predict outcome of election by collecting twitter data and analyze it to predict the outcome of the TELANGANA election by analyzing sentiment of twitter data about the parties and candidates. For this, the idea is to combine the data mining technology using machine learning approach to find emotions in twits and predict sentiment score.

Keywords: *Sentiment analysis, Machine Learning, Natural Language Processing, Python, Election Prediction*

1. Introduction

An analysis of the tweets' political sentiment demonstrates close correspondence to the parties' and politicians' political positions indicating that the content of Twitter messages plausibly reflects the offline political landscape. The use of microblogging message content as a valid indicator of political sentiment and derive suggestions for further research. The aim of this study is three fold

1. Examine whether Twitter is a vehicle for online political deliberation by looking at how people use microblogging to exchange information about political issues.
2. Evaluate whether Twitter messages reflect the current offline political sentiment in a meaningful way.
3. Analyse whether the activity on Twitter can be used to predict the popularity of parties or coalitions in the real world. Lately, it is observed that traditional polls may fail to make an accurate prediction.

The scientific community has turned its interest in analysing web data, such as blog posts or social networks' users' activity as an alternative way to predict election outcomes, hopefully more accurate. Furthermore, traditional polls are too costly, while online information is easy to obtain and freely available. This is an interesting research area that combines politics and social media which both concern today's society. It is interesting to employ technology to solve modern-day challenges.

2. Related Work

This paper mainly focuses on defining related work about predicting the TELANGANA election based on twitter data and it can also analyzing the political sentiment all over the India. In current, Social media is a relatively new form of

communication, analyzing web data and casting predictions based on that data is a popular subject for research. It is to be three research areas emerging in terms of using online sentiment to monitor real world political sentiment.

First, is event monitoring, where the aim is to monitor reactionary content in social media during a specified event. In the political area this would typically be a speech, TV debate, In Foreign Country Twitter proved to be an effective source of data for identifying important topics and associated public reaction.

Second, is using selected data prior to the election. The prediction could be derived by comparing the number of tweets mentioning each candidate or by comparing the number of tweets that has positive sentiments towards each candidate. The earliest research stated that the number of tweets mentioning a party reflects the election result where they found out that the prediction result from Twitter were only better than other.

Third, related area is result forecasting. In result forecasting, the result which is used to judge the accuracy of a particular forecasting measure, rather than a continuous series. Asur and Huberman (Asur and Huberman, 2010) used Twitter volume and sentiment to predict box office takings for movies, bettering other market indicators. It finds volume be a strong predictor and sentiment to be a useful, yet weaker predictor. They also propose a general model for linear regression social media prediction which serves as a basis for model. Thus seeing the predictive systems which utilize social media are both promising and challenging. The contention of research is that the development of techniques for political public sentiment analysis and election prediction is a promising direction requires more research work before understands the limitations and capabilities.

3. Methodology

Initially, collect the political data and then pre-process each party's tweets in one of four ways: (i) general mentions with no filtering, (ii) specific mentions or retaining only those tweets that mentioned a single party at a time, (iii) positive tweets or retaining only those specific mention tweets that had positive sentiment, and (iv) temporally weighing tweets that were closer to the election. Then compare the different pre-processing methods and approaches in terms of their ability to predict the actual vote share of each party. These methods are detailed in the following sections.

3.1. Data Collection

The data collection step is the initial phase in the research, where data is collected from twitter. There are two methods on how to connect and collect tweets from Twitter. The first method is by searching tweets matching to the keywords. The second method is by collecting all the tweets provided by Twitter through streaming API (Application Programming Interface), or all the tweets in a specific language, or all the tweets in a specific location then put all of them into the database.

API is a way for software to access the Twitter platform (as opposed to the Twitter website, which is how humans access Twitter). While supporting a large number of functions for interacting with Twitter, the API functions most relevant for acquiring a Twitter dataset include:

- Retrieving tweets from a user timeline (i.e., the list of tweets posted by an account Searching tweets.

- Filtering real-time tweets (i.e., the tweets as they are passing through the Twitter platform upon posting).

3.2. Data Pre Processing

An initial step in text and sentiment classification is pre-processing. A significant amount of techniques is applied to data in order to reduce the noise of text, reduce dimensionality, and assist in the improvement of classification effectiveness. One of the most important goals of pre-processing is to enhance the quality of the data by removing noise. The most popular techniques include: -

i) Lower Case Conversion: Because of the many ways people can write the same things down, character data can be difficult to process. String matching is another important criterion of feature selection. For accurate string matching are converting the complete text into lower case.

ii) Removing Numbers and Removing Punctuation's: All punctuation's, numbers are also need to remove from reviews to make data clean and neat. Unnecessary commas, question marks, other special symbols get removed in this case. Here, not removing dot (.) symbol because it's splitting our text into sentences.

iii) Stemming: Stemming is that the method of conflating the variant styles of a word into a standard illustration, the stem. For example, the words: "presentation", "presented", "presenting" could all be reduced to a common representation "present". This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented.

iv) Striping White Spaces: In this pre-processing step all text data is cleansed off. All unnecessary white spaces, tabs, newline character get removed from the text.

3.3. Feature Extraction

Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. This can also be used to enhance the speed and effectiveness of supervised learning. Extraction of ten to twelve features and categories as Tag based features and URL based features. User-based features were extracted from the JSON object "user," User-based features, like no of followers, no_of followings, no_userfavourites, no lists, and no tweets, can be directly parsed from the JSON structure. Tweet-based features include no_ retweets, no_ hashtags, no_usermentions, no_urls, no_chars, and no_digits. While no_chars and no_digits need a little computing, i.e., counting them from the tweet text, others can also be straightforwardly extracted.

3.4. Data Classification

In classification, the idea is to predict the target class by analysis the training dataset. This could be done by finding proper boundaries for each target class. In a general way, Use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions determined, the next task is to predict the target class. In this, the whole process is known as classification.

In this evaluated the solution with different machine learning algorithms namely Support Vector Machine, Bagging Algorithm (Bootstrap Aggregation). With SVM, are

able to achieve accuracy. In the proposed framework, the ML-based predictor receives as input a feature vector, which characterizes the incoming tweet according to a feature space. Finally, predict politicians campaigning and strategy and who will win or comes to nearest based on getting a positive comment.

4. Sentiment analysis

Sentiment analysis in machine learning is a process of automatically identifying whether a user-generated text expresses positive, negative or neutral opinion about an entity (i.e. product, people, topic, event etc).ML approach classified into two ways

1. Supervised
2. Unsupervised.

i) Supervised Machine Learning

In this paper supervised machine learning is used where input variables (Tweets from different users like x) and an output variable (predicted output like Y) and algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
Supervised learning problems can be further grouped into regression and classification problems.

- A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

ii) Unsupervised Machine Learning

Other is unsupervised learning is where only have input data (like X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. These are called unsupervised learning because unlike supervised learning there is no correct answer and there is no teacher.

Unsupervised learning problems can be further grouped into clustering and association problems.

- A clustering problem is where discover the inherent groupings in the data, such as grouping tweets of particular election candidate.
- An association rule learning problem is where discover rules that describe large portions of data, such as people that buy X also tend to buy Y.

5. Classification Algorithm

ii) Support Vector Machine (SVM)

Support vector machine is non probabilistic algorithm which is used to separate data linearly and nonlinearly. It’s determines the best decision boundary between vectors that belong to a given group (or category) and vectors that do not belong to it. That’s it. It can be applied to any kind of vectors which encode any kind of data. This means that in order to leverage the power of SVM text classification, texts have to be transformed into

vectors. Vectors are (sometimes huge) lists of numbers which represent a set of coordinates in some space.

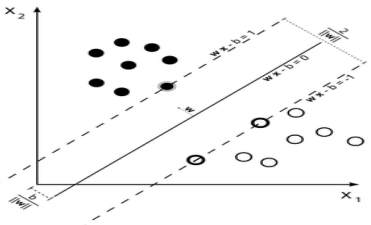

Figure-1

In the figure 1 point plotted in 2D-space. The set of points are labelled with two categories (illustrated here with black and white points) and SVM chooses the hype plane that maximizes the margin between the two classes. This hyper plane is given by

$$\langle \vec{w} \cdot \vec{x} \rangle + b = \sum_i y_i \alpha_i \langle \vec{x_i} \cdot \vec{x} \rangle + b = 0$$

Where xi= (xi1, xi2,..., xin) is a n-dimensional input vector, yi is its output value w= (w1,w2,...,wn), is the weight vector (the normal vector) defining the hyper plane and the terms are the Lagrangian multipliers.
Once the hyper plane is constructed (the vector is defined) with a training set, the class of any other input vector can be determined:
if w.xi+b>=0 then it belongs to the positive class (the required class), otherwise it belongs to the negative class (all of the other classes).

ii) Bagging (Bootstrap Aggregation)

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

To improve the limited classification performance of SVM, use the ensembles with bagging (bootstrap aggregating). Each individual SVM is trained independently using the randomly chosen training samples via a bootstrap technique. Then, they are aggregated into to make a collective decision in several ways such as the majority voting, the LSE (least squares estimation)-based weighting, and the double-layer hierarchical combining. Various simulation results for the data classification show that the proposed SVM ensembles with bagging outperform a single SVM in terms of classification accuracy greatly.

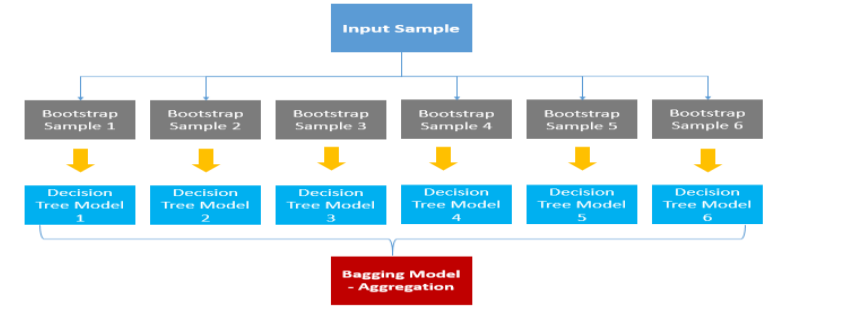

Figure-2

4. Result Analysis

The data collected through twitter API, after that performed pre-processing steps on that data. The output results are Classifying in three categories positive, negative and neutral based on the tweets.

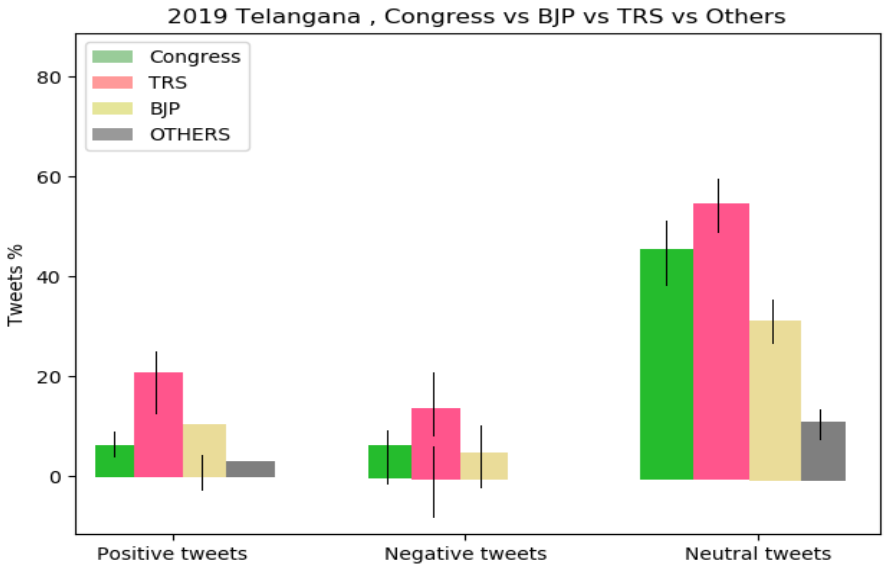

Figure-3 overall TELANGANA state report 2019

The above figure illustrates the overall TELANGANA state report where four parties is CONGRESS, TRS, BJP, OTHERS (remaining parties) taken. The level of sentiment analysis filters out twitter sentiment which is determines whether the customer tweets about the parties is positive, negative and neutral in the above figure.

```
Python 3.6.7 (v3.6.7:6ec5cf24b7, Oct 20 2018, 13:35:33) [MSC v.1900 64 bit (AMD6
4)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
== RESTART: C:\Users\Kishan\Desktop\TELANGANA ELECTION\twitter_election.py ==
Enter Keyword/Tag to search about: KCR as CM
Enter how many tweets to search: 2000
How people are reacting on KCR as CM by analyzing 2000 tweets.

General Report:
Weakly Positive

Detailed Report:
0.15% people thought it was positive
0.05% people thought it was weakly positive
0.00% people thought it was strongly positive
0.05% people thought it was negative
0.00% people thought it was weakly negative
0.00% people thought it was strongly negative
0.45% people thought it was neutral
```

Figure-4 Text report by Entered Keyword

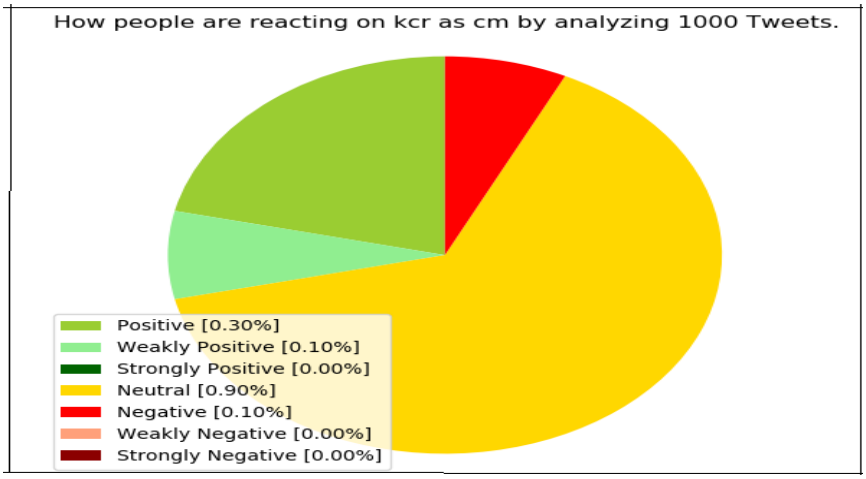

Figure-4.1 Graph report by Entered Keyword

In the above, Figure 4 and 4.1 shows the result based on "searched keyword" in this the search keyword is "KCR as CM" and the result is classified based on positive, negative and neutral tweets.

5. Conclusion

In this paper, twitter data, where the discussion about the election is started to be posted, until the time conducted the experimental work. Based on that data, a new method is used to predict the election result that focuses only on tweet counting and sentiment analysis as the pre-processing task. In this access the tweets of candidates using Twitter API. This method is a simpler than other methods yet proved to be sufficient to produce a reliable result since both aspects have a significant contribution to the prediction. The experimental result is produced by using Python language. This prediction result is corresponding to the TELANGANA state election in India. For the future works, this paper will continue mining and analyzing more Twitter data until around the election time and after the election to get a more accurate prediction.

6. *References*

[1] *Sentiment Analysis of Election Result based on Twitter Data using R Volume: 05 Issue: 03 | Mar- 2018*

[2] *Twitter Based Election Prediction and Analysis Volume:04 Issue: 10|Oct-2017*

[3] *Hamling T, Agrawal A. Sentiment analysis of tweets to gain insights into the 2016 US election. Columbia Undergraduate Sci J. 2017;11:34–42.*

[4] *Makazhanov, A. R. (2014). Predicting political preference of Twitter users. Social Network Analysis and Mining, 1-15S.L. Mewada, "Exploration of Efficient Symmetric AES Algorithm", International Journa of Computer Sciences and Engineering, Vol.4, Issue.11,pp.111-117, 2015.*

[5] *Boutet, A. K. (2012). What's in your Tweets? I know who you supported in the UK 2010 general election. Proceedings of the International AAAI Conference on Weblogs and Social Media.*

[6] *Sang, E. T. (2012). Predicting the 2011 dutch senate election results with twitter. the Workshop on Semantic Analysis in Social Media (pp. 53-60).Association for Computational Linguistics.*

[7] *Fumagalli,L.&.(2011).The total survey error paradigm and pre-election polls: The case of the 2006 Italian generalelections. ISER Working Paper Series2011.*

[8] *Hillygus,D.S.(2011).The evolution of election polling in the United States. Public opinion quarterly,75(5),962- 981*

[9] *Pak, A. &. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining.LREC.Dann, S. (2010). Twitter content classification. First Monday,15*

[10] *LewisBeck,M.S.(2005).Election forecasting: principles and practice. The British Journal of Politics & International Relations, 7(2),145-164.*

Author Biography

Kishan Kumar Sahu *pursuing Master of Computer Applications, GIs, GITAM (Deemed to be University, Visakhapatnam. His area of interest in Machine learning tools and algorithms.*

K Vanitha *is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University). Her main areas of research include Data Mining and image processing.*