International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Opinion Mining Using Twitter Feeds for Political Analysis

Manogna Meduru[a*], Antara Mahimkar[b], Krishna Subramanian[c], Puja Y. Padiya[d], Prathmesh N. Gunjgur[e]

[a,b,c,d,e]*Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India*

[a]*Email: mmanogna29@gmail.com*

[b]*Email: antara216@gmail.com*

[c]*Email: krishnasubramanian16@gmail.com*

[d]*Email: puja.padiya@rait.ac.in*

[e]*Email: prathmesh.gunjgur@rait.ac.in*

Abstract

Sentiment analysis deals with identifying and understanding opinions and sentiments expressed in a particular text. The masses give their opinion regarding various subjects on social media platforms using tweets, status updates and blogs. By analyzing this very data, we can gain better insight of the public opinion on any subject in specific. On performing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters allowed in Twitter is 140. In this paper, we try to analyze the twitter posts about government issues and political reforms. The proposed framework uses Twitter as the platform to analyze the emotions of the users using Sentiment Analysis. The system will use the opinions of the users, analyze the reaction and then map it to the appropriate region.

Keywords: Opinion mining/Sentiment Analysis; Natural language Toolkit(NLTK); Political Analysis.

1. Introduction

In present world scenario, we are dealing with a huge and an unprecedented amount of opinionated data that is available on the social media. The field of data analytics would not exist if it were devoid of such data. Sentiment analysis, also known as opinion mining, is the branch of data analytics, which deals with analyzing an individual's opinion and emotions towards particular services, related organizations, trending issues and events. Microblogging sites have become a huge mine of information on various subjects.

---

* Corresponding author.

Owing to the nature of these websites, people actively post content on currently trending topics giving their frank opinion. These messages give interesting patterns and serve as a great data source for sentiment collection. The technology has reached its pinnacle since the birth of AI. Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI), which helps in understanding the colloquial language used by human in order to interact with computers, both in written and spoken context.

Here, we look at one such popular microblog called Twitter and build models for classifying "tweets" into positive, negative and neutral sentiment. Twitter is one of the most concise and precise microblogging sites. Individuals always tweet over latest happenings with a "Hashtag" which is frequently the fundamental subject for examination. The data is collected using the streaming API of twitter. One advantage of this data over manually used data sets is that the tweets are collected in a streaming fashion and therefore represent a true sample of actual tweets in terms of language use and content. Sentiment Analysis on a sentence-level can be done using lexical approach. Vader, an open source tool and a part of NLTK python library, carries out the part of the sentiment analysis on a sentence-level. A method named SentimentIntensityAnalyser under Vader is useful to calculate the polarity of the text and classify them accordingly among positive, negative or neutral.

2. Literature Survey

Sentiment Analysis is the computational investigation of individual's conclusions, notions, demeanors, and feelings communicated in composing dialect. It is a standout amongst the most dynamic research ranges in regular dialect handling and machine learning lately. Its prominence is for the most part because of two reasons. To start with, it has an extended variety of utilizations since feelings are central to all human practices and are key influencers of our patterns. At whatever stage we have to go down on a choice, we necessitate to hear other's conclusions.

RaviChandran *and his colleagues* [1] has dealt with the issue of identifying extremity of words by and recognized about the extremity of a word to either be certain or negative. For instance, words, for example, great and brilliant, are considered as positive words; while words, for example, awful, malice, and evil are viewed as negative words. The polarity detection is treated as a semi-supervised label propagation problem in a graph.

Delenn Chin *and his colleagues* [2] concentrated on tweets with respect to the presidential elections. They just considered tweets with emojis relevant to feelings about the candidates and then categorized them by creating a mapping of emojis to sentiments. Their key feature was the construction of maps from the sentiments of the tweets. We discovered that for accessing the location of a tweet, it is required that the user enables this setting to show longitude and latitude coordinates.

Sentiment analysis for sentence-level can be measured using the lexical approach. It conveys the polarity information. Filipe N Ribeiro *and his colleagues* [3] performed an analysis of 24 various sentiment analysis tools. The authors postulate that the selection of sentiment analysis approach is eminently dependent upon the type of application. The methods, which fared well in their experiment included VADER, SentiStrength, etc. amongst which VADER, an open source tool and a part of the NLTK python library, was concluded to be the

most efficacious for social network text platforms like Twitter.

C.J. Hutto *and his colleagues* [4] concentrated on analyzing a corpus by comparing different sentiment analysis lexicons. It was observed that VADER outperformed the other established lexicons, especially well in the social media domain. They compared its effectiveness to machine learning techniques like SVM. We learnt that SVM takes longer than VADER to analyze a large corpus. A few researchers like Bo Yuan [5] publish better results for machine learning-based methods accompanied by some restrictions.

Sentiment analysis can be performed using either lexical-based methods or machine learning-based approach. The supervised machine learning-based methods have the ability to adapt and create trained models for specific purposes and contexts. In the aforementioned papers, there is the need of labeled data in machine learning-based methods, which might be highly costly, or even prohibitive, for some tasks. On the other hand, the lexical-based methods make use of a pre-defined list of words, where each word is associated with a specific sentiment – positive, negative and neutral. Lexicon provides sentiment information about the smallest linguistic unit.

Vader is a simple lexicon-based and rule-based approach, which performs across social media domain**.** It utilizes a general lexicon and rules making it free of the requirement of a training data set**.** It makes the interior work of the sentiment analyzer engine more accessible, while pertaining to its accuracy and speed, making it the best option for measuring sentiment polarity.

2.1 Similar Existing System

Andranik Tumasjan and his colleagues [6] proposed a system for the predication of election results in Germany that aims at twitter examination for a political vehicle, to evaluate tweets reflection of current political sentiment. A huge data set was analyzed using LIWC to extract sentiment. On analysis, it was found that unprecedented words were predicted incorrectly. The use of VADER has proved to be beneficial as VADER, unlike LIWC, is sensitive to sentiment expressions on social media and therefore improves the accuracy of categorizing the sentiment.

Preslav Nakov and his colleagues [7] proposed a system that uses SentiWordNet 3.0 to extract and classify the sentiment from the tweets. With an addition in their training dataset, the testing data's categorizing accuracy (as positive, negative and neutral) is improved. Lexicon approach proves to reduce the trouble of updating the training data set every time a new word is encountered. In addition, it uses a sentence level approach to find the sentiment of the message as opposed to finding sentiment of each word. Multiple lexicon approaches are available and VADER proves to be the best for our system.

From the study of the existing systems, we infer that there is a need for a better system that handles all the ambiguities to improve the analysis part. Supervised learning algorithms require a training data set, which needs to be updated every time a new word is encountered. In addition, algorithms like Naive Bayes and SVM do not prove to be useful when a huge data set is involved. Unsupervised algorithms serve the purpose in case of huge dataset. The unsupervised algorithms avoid the trouble of clustering and directly categorize them to their respective sentiment category (namely positive, negative, and neutral) based on some predefined values or rules.

Lexicon analysis helps in identifying the sentiment on a sentence level rather than on individual level.

VADER and LIWC are two useful approaches for this purpose. However, VADER is preferred because it is more sensitive to social media sentiment than LIWC. A detailed approach is mentioned in the proposed methodology.

3. Problem Definition

Opinion is subjective to a person and varies along with a lot of parameters. Opinion referred to as sentiment sometimes has some close association with region and such other distinguishing features. Politics being a matter of public interest and one that catches curiosity. It is one of the most important matters in every country. It can prove to be a good source of information and can be of use to the public, if the way the public perceives the changes that happen around is analysed. Keeping that in mind we have decided to create a system that uses data mining approach to find patterns in this data and predicts the sentiment of the data gathered and gives a region wise classification.

These reforms are implemented for the betterment of the public. These public, in the modern digital world, can express their opinion or Sentiment online as soon as the speculated policy or the policy yet to implement comes under the limelight. The proposed system uses twitter as the platform to collect data in order to analyse the sentiment of the common public towards the policy. The application gathers tweets on a particular hashtag (user entered custom search field) and their respective location so that the sentiment can be mapped to a location. The sentiment can be categorized as either positive, negative or neutral based on certain keywords. The system fetches the location and analyses the sentiment of the user and represents the overall reaction from a particular region (North, South, East or West) graphically.

4. Proposed Methodology

Opinion Mining Using Twitter Feeds for Political Analysis helps in understanding the responses of the masses with respect to the governmental issues and political reforms.

Our system is an application based system where in the user has to enter a suitable hashtag, the one for which he intends to perform the analysis and obtain the results. The user has to type in the hashtag along with the number of live tweets one wishes to analyze (with a limit of 2000 tweets at a particular time).

Once done with the aforementioned stipulations, the user has to click on results tab to generate the results. The system uses the dataset obtained from the live tweets and performs the sentiment analysis to identify the general reaction.

The detailed functioning of the application is explained as follows:

4.1 Data Collection

The data collected using the Streaming API of Twitter as Live Tweets is taken into consideration.

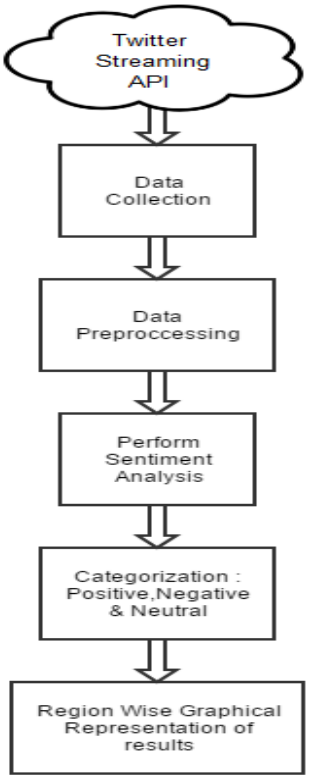

Figure 1: Architecture of the System

4.2 Text Pre-processing

Pre-processing of data is the process of preparing and cleaning the data of dataset for evaluation of the sentiment value. The reduction of the noise, in the text, should help improve the performance of the system and speed up the analysis process, thus aiding in real time sentiment analysis.

1. Stop word Removal: A stop-list is the name commonly given to a set of stop words, which do not contribute to the meaning of the sentence. Some of the more frequently used stop words for English include “a”, “of”, “the”, “I”, “it”, “you”, and “and”. The meaning can be conveyed more clearly by ignoring the functional words and hence, it is practical to remove those words.
2. Slang Removal: Due to the 140-character limit in Twitter, the users tend to make use of slangs as opposed to the normal diction. However, these slangs, according to the English dictionary, are incorrect use of the language. Hence, in this step, the slangs are replaced with their appropriate form.
3. URL, Hashtag and Emoji Removal: Twitter users tend to include URLs in their tweets with the employment of hashtags (relevant to the subject) and use emojis, which do not carry any inherent meaning within themselves with respect to this application.
4. Special characters: Characters like “@”, “$”, “%”, etc. do not have any relevance as far as the meaning is concerned.

4.3 Perform Sentiment Analysis

After refining the live data stream, the next task is to measure the sentiment polarity information. For this, the Vader lexicon is used which categorizes the sentences based on a predetermined lexicon and some rules for analysis. It compounds the values of all the words, calculates the overall sentiment, and divides it amongst either of the following:

1. Positive
2. Negative
3. Neutral

4.4 Region-wise Graphical Representation

Once the entire sentence's polarity is calculated, the same is mapped with its respective region captured during the tweet collections. A pie chart with three different color codes, one for each category, is shown for every region. Based on the responses of the public, the user can act accordingly and enhance the policies if necessary.

5. Result Analysis

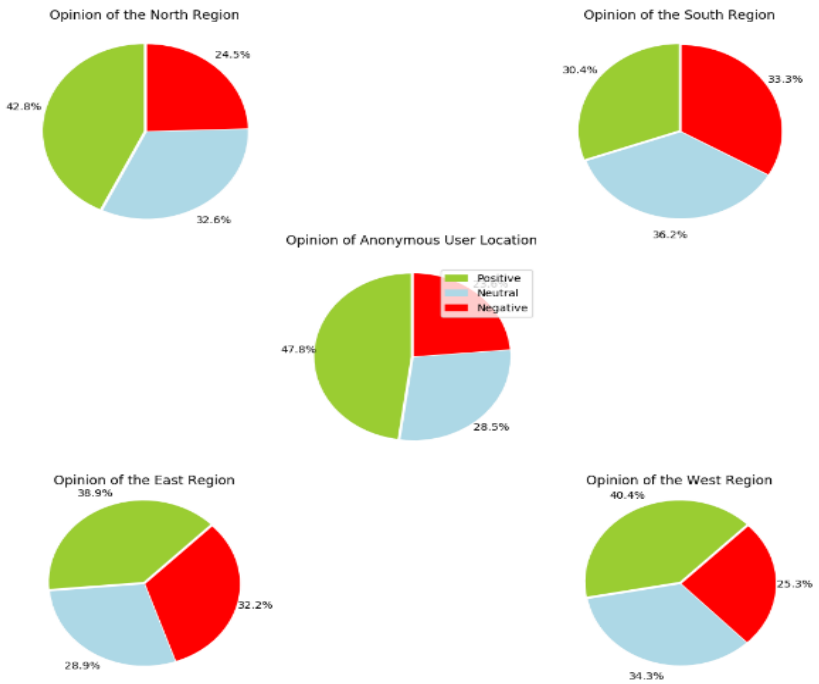

Figure 2: Result of the System

The above graphs represent the response to the hashtag "Modi". It represents an integration of the masses' opinion from North, South, East and West of India along with the users who have disabled their location on Twitter. The regions in Red represent 'Negative' sentiment; Green regions represent 'Positive' sentiment while Blue regions represent 'Neutral' sentiment.

Table 1: Data collection and analysis for the result

	POSITIVE	NEGATIVE	NEUTRAL	COUNT
NORTH	311	178	237	726
EAST	35	29	26	90
SOUTH	95	104	113	312
WEST	112	70	95	277
DISABLED	238	140	169	547
TOTAL	791	521	640	1952

The above table helps us to understand exactly how many tweets were collected and what sentiment was found in the respective regions. Some users who have disabled their location are taken into consideration in the disabled sections. These statistics were used to plot the graphs for better understanding.

6. Limitations of the System

VADER, although very efficient, is a human validated sentiment analysis method therefore is prone to some issues. It has a lexicon size of merely only 7517. The proposed system analyses twitter feeds wherein slang is commonly used, the internet vocabulary is ever increasing, VADER's lexicon size may seem insufficient in the near future and provide lesser accuracy. Sarcastic comments are difficult to identify and can be classified as neutral or negative because of the tone in the absence of accurate emoji.

7. Conclusion

Opinion mining is the process of extracting the sentiment from the text or a given sentence. The use of this tool in political analysis is relatively new and unexplored. People and the government feel a lack of connection between them and directly or indirectly, the failure of any policy is a result of this communication gap. Our system reduces the gap and make the otherwise inactive participation between the two sides, an active one. The proposed system uses Twitter as the platform to analyze the emotions of the users using Sentiment Analysis. The system uses Vader Lexicon to determine probabilistic sentiment of a tweet.

References

[1] Delip Rao, Deepak Ravichandran. (2009). Semi-Supervised Polarity Lexicon Induction.

[2] Delenn Chin, Anna Zappone, Jessica Zhao. Analyzing Twitter Sentiment of the 2016 Presidential Candidates.

[3] Filipe N Ribeiro, Matheus Ara´ujo, Pollyanna Gonc¸alves, Marcos Andr´e Gonc¸alves, Fabr´ıcio Benevenuto. (2016, July). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods.

[4] C.J. Hutto, Eric Gilbert. (2014).VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.

[5] Bo Yuan. (2016). Sentiment Analysis of Twitter Data.

[6] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.

[7] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani|, Veselin Stoyanov. (2016, June). SemEval-2016 Task 4: Sentiment Analysis in Twitter.

[8] Pang, B., Lee, L. & Vaithyanathan, S. (2002). Sentiment Classification using Machine Learning Techniques.

[9] Gebrekirstos Gebremeskel. (2011, May). Sentiment Analysis of Twitter Posts about News.

[10] Effective Text Data Cleaning [Online]. Available

https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/

[11] Mining and Preprocessing Twitter Data with Python [Online]. Available

https://marcobonzanini.com/2015/03/09/mining-twitter-data-with-python-part-2/

[12] Useful Pandas Techniques in Python for Data Manipulation [Online].Available https://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data-manipulation/