International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Real Time Sentiment Analysis of Political Twitter Data Using Machine Learning Approach

Joylin Priya Pinto[1], Vijaya Murari T.[2]

[1]*Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India*
[2]Professor, *Department of Computer Science and Engineering, NMAMIT, Nitte, Karnataka, India*

Abstract - *Social media is an interesting platform to directly measure people's feelings. Communication technologies play a vital role in geographically locating these emotions. But, the understanding is not a simple task. Generation of voluminous data makes the manual process complicated. Social media data face the problem of diversity of language; due to which the automatic approach also becomes a tedious job. Social network deals with controversial discussions on variety of topics. These discussions help in the field of data analysis. In this paper, we are going to analyze tweets on political Ayodhya issue. Tweets of users are collected and analysis is done with the help of machine learning algorithm, to classify the polarity of tweets.*

Key Words: **Sentiment Analysis, Feature Extraction, Support Vector Machine, Random Forest, Naïve Bayes, Linear Regression, KNN**

1.INTRODUCTION

Analyzing the sentiments of people is turning into a critical viewpoint in a wide range of decision making process since it is useful in recognizing individual's issues and strategies strengths. Without a doubt, these information can be utilized to settle on increasingly educated choices which will probably conclude in better utilization of assets, good association, better administration, improved citizen lifestyle, nice human relations and, in the long run, better society. Taking are of large scale classification issues is very essential in numerous applications, for example, text classification problem. There is a high possibility of facing distinctive troubles while performing sentiment analysis; for each situation people may not present their assumptions similarly, a solitary sentence may be certain for one event and can be negative for other event; an enormous number of sentence mixes are possible. Identifying incorrect spelling, and dealing with intensifiers and fake sentences are very challenging.

In the past few years, people feedback and sentiments were analyzed by opinion pooling system that means with the help of conducting interviews, questionnaires and by collecting opinions through forms, [1][2] now a days social media is the best way to analyze people sentiments [3]. There is an immense growth of user-generated data in the form of blogs, forum and tweets. With the increased usage of these platforms, people started to speak about all matters: from personal to public; general to specific matters. Social media is an effective platform to understand individual sentiments. The analysis of user posts can be utilized to take proper decisions in a variety of fields such as Business, Election, Product review, Government, and so on. Now a days, utilizing social media for political discussions has become a common practice. Political campaigns have misused immense range of information accessible on Twitter to draw insights about people sentiments and in this way structure their promoting efforts. Different sentiment analysis algorithms can be utilized to distinguish and break down the attitudes of the users towards a political talk. Furthermore, with the recent advancements in machine learning algorithms, it became possible to enhance the exactness of sentiment analysis predictions.

An effective online social micro-blogging service is Twitter. It allows users to communicate with short messages of length 140-characters. These messages are called as "tweets". Twitter is specially an interesting platform because of the usage of its hash tags. Along with the short messages, users can use hash tag symbol '#' before a specific keyword or phrase in the Tweet to arrange the tweets and make them easier in Twitter Search. The issue of text classification can be made relatively simpler since the hash tag itself can express a feeling or sentiment. In this work, twitter data on Ayodhya issue is used as data source for sentiment analysis.

1.1 Sentiment Classification

Sentiment classification is one of the important topics in sentiment analysis, which classifies the expressed opinion towards an entity as positive, negative or unbiased. Three most essential categorical levels in Sentiment Analysis are Document level, Sentence level, and Feature based Analysis [4]. The underlying stage in Sentiment Analysis is to perceive whether the sentence is opinionated or objective. Subjective sentence conveys emotions whereas objective sentence does not convey any emotion; usually considered as unbiased and also being ignored in the analysis. Sentence level classification is different from that at document level [5]. A document can be more or less opinionated, whereas a sentence can be only subjective or objective. Comparing with the document level sentiment classification, sentence level classification has one more task to do. There is a need to

process the sentences containing no opinion with sentences containing opinions before classifying on the objects and their features to positive or negative.

1.2 Entity Discovery and Extraction

People communicate their opinions and sentiments against an entity or object. Sentiment analysis and opinion mining can be formalized as: "Given a set of evaluative text documents D that contain sentiments about an object, sentiment analysis and opinion mining aim to extract attributes and components of the object that have been commented on in each document d in D and to determine whether the comments are positive, negative or neutral." To perform sentiment analysis, along with distinguishing between polarities of opinions, identifying correlated entities with which opinions are expressed is also very important. Because without analyzing about which entity each sentence talks about, the mined opinion is meaningless.

2. LITERATURE REVIEW

Many researchers have done lot of work on "Sentiment analysis" in past few years. Since the starting of the century itself the work in the field had begun. In the beginning period, sentiment analysis was intended for binary classification, where opinions were assigned to bipolar classes as positive or negative.

Extensive consideration has been given as of now to the exploitation of big datasets generated by the users through social media platforms with which social behavior of the people can be identified. In order to naturally recognize people sentiments, few approaches such as mentioned in [6] use linguistic analysis systems; some depend on manual approach [7] and some other depends on a self-loader approach [8]. All the methods have pros and cons. The manual approach provides more accurate result. But, it does not work well while dealing with huge volumes of data; on the other hand, an automatic approach is difficult to implement. Because it has to deal with natural language processing; there is a big challenge to work with the characteristics of the posted content. For example, if tweets are considered, they are normally posted with hashtags, emoji's and url links, which makes it difficult to determine the expressed sentiment [9]. Moreover, training is needed to automatic process. Training requires large dataset of labelled posts or lexical database where opinionated words are labelled with sentiment values. These resources are available for the English language [8], [9], [10], but the availability of dictionary based dataset is very limited to work with other languages [7], [11].

A. khan et.al [12] used an approach to compare between positive and negative sentences. It extracts information from the Web and manually label the word set which requires a lot of unnecessary effort. He has used a rule-based method for sentiment analysis, which extract the overall document polarity of specific words by a SentiWordNet dictionary, and adjust it according to the context information. First, sentences are divided into subjective and objective ones on the lexical dictionary basis. The subjective sentences are then further processed to categorize as positive, negative or neutral comments.

Judith et.al [13] focused on twitter data analysis techniques to extract public opinion. Using machine learning algorithm, a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion. It said that various twitter data analysis techniques are based on dictionary and are using the machine learning approaches.

Barbosa et. al. [14] designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to reduce the labelling effort in developing classifiers. Firstly, they classified tweets into subjective and objective tweets. After that, subjective tweets are classified as positive and negative tweets.

Pak et. al. [15] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that method, there is a chance of error since emotions of tweets in training set are labelled solely based on the polarity of emoticons. The training set is also less efficient since it contains only tweets having emoticons.

Upma Kumari et. Al. [16] carried out a detailed study on the sentimental analysis techniques and tools which have been used in sentiment analysis process. She observed the complete process that shows how the input is being classified in various stages of sentiment analysis. According to her, first step of text summarization process is sentimental information retrieval, then classification and finally sentiment summarization process. Subjectivity of text is checked first and based on the polarity, sentiment results are generated. The proposed method provides important phases of sentiment analysis on text classification.

Seyed-Ali et. al. [17] proposed a novel approach to sentiment analysis of short informal texts such as tweets posted in twitter. He used state-of-art Sentiment analysis technique against a novel hybrid approach. The method uses a sentiment lexicon to generate a set of new features to train a SVM classifier. Author focused on the processing of twitter data. Then tried to apply Support Vector Machine, Naïve Bayes as well as Maximum Entropy classifiers. Accuracy of the model is compared against SVM and hybrid approach and author proved that the novel hybrid method which he proposed is more accurate than the support vector machine model. Also the features presented in this approach does not require more time to compute the result which is better than other techniques.

Devika et. al. [18] proposed her views on sentiment analysis as it is a process of human feelings and opinion extraction. It poses as the most powerful tool to get useful

information from the users which can be then used to aggregate the collected sentiments from reviews. She has discussed different techniques of sentiment analysis by the analysis of various methods. Machine learning algorithms such as SVM, N-gram, Naïve Bayes, KNN, Feature driven sentiment analysis as well as rule and lexicon based approaches are focused. The purpose of the paper was to come out with a best technique to text classification approach.

3. RELATED WORK

3.1 Conceptual Sentiment Analysis Framework

Twitter Sentiment Analysis Model does the analysis of sentiments as an automatic process. It identifies the opinions as positive or negative and measures the strength of expressed opinion on a topic or an entity. The conceptual model of Twitter Sentiment Analysis has the following modules:

- ✓ Feature Selection Module which does the opinion extraction job. It extract the words that express opinions towards relevant entity at sentence level analysis.
- ✓ Sentiment Detection Module which maps the expressed opinion with respect to relevant entity in each sentence.
- ✓ Sentiment Summation and Score Calculation Module to measure the sentiment scores for each entity.

Fig. 1 illustrates the conceptual framework of Twitter Sentiment Analysis. The description of each module is stated below.

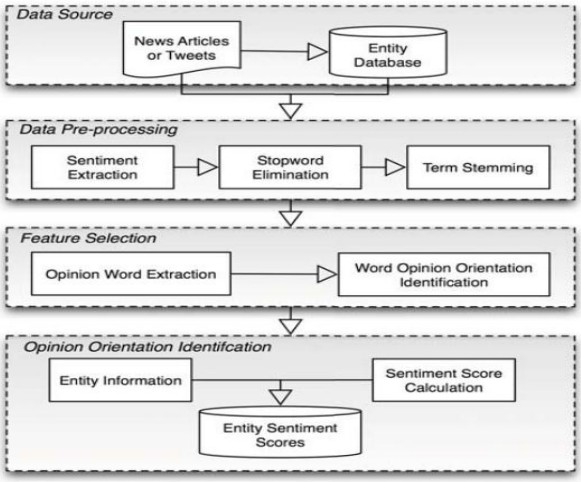

Fig -1:The Conceptual Sentiment Analysis Framework [20]

Data Collection: In the data collection process, data from an intended source is collected in a systematic way. Then the collected information is processed to produce better outcome. Here, this sentiment analysis study is especially for political domain and for the political related analysis, political based comments are required; political tweets can be tweets related to discussion between political parties, about famous political leader, discussions on upcoming elections or any trending political issues. Twitter allows to gather data for the academic or research purposes. Relevant data can be collected by providing search keywords in the twitter search field. In order to obtain twitter data, there is a need to create Twitter developer account. Twitter requires the user to be registered with the developer account and once the registration is done, it provides necessary authentication credentials when user queries the Twitter API. The user account need to be verified and only if the account is valid, twitter allows the user to access the API.

Here, the analysis is mainly on 'Ayodhya issue'. The tweets are scraped from the Twitter with the help of user credentials. As there is no availability of polarity based Indian political training dataset, a new training dataset is built by using Vader Sentiment Analyzer. Initially, training is done based on this dataset for the collected twitter data. After testing, the predicted tweets are added to the training set and newly scraped tweets are kept for testing purpose.

Data Pre-processing Module: Text pre-processing is the stage, where the raw data is transformed into something that a classifier can read. User-generated data may contain URL links, unnecessary noisy data. Pre-processing cleans the data by removing URL's, user names, stop-words and all irrelevant data that does not express any sentiment. Here, the research work uses hash tags and emoticons in the analysis of sentiments in order to improve the efficiency.

In the sentiment analysis process, text pre-processing plays vital role. No doubt, qualitative data increases the efficiency where as an incomplete and noisy data reduces the accuracy and effectiveness.

Feature Extraction: Most of the times sentiment analysis relies on "Bag-of-Words" model, which is the representation of text that describes the occurrence of words within a sentence or document [19]. This is a very common feature extraction method. Here, in this sentiment analysis model, only the sentiment bearing words are extracted as features in order to submit to the classification algorithm; remaining words are ignored. Opinionated words usually express subjective opinions. Here, in the feature extraction process, words which exhibits a desirable state (e.g. good, wonderful) express positive orientation whereas the words with undesirable state shows a negative orientation. (e. g. very sad)

Sentiment Analysis Method: Consider a set of tweets 'T', which has a set of sentences S = {s1, s2, ... si} and each sentence st express some opinions on an entity e. An entity could be a person, a place, a topic etc. In this study, entity is a topic i.e. "Ayodhya" issue. Each sentence s1 may contain set of opinionated words such as s1 = {w1, w2, ... wi}. Firstly, sentiment score of each sentence will be calculated based on expressed words. Secondly, sentiment summation will be

done to obtain the total sentiment scores towards entity e. [20]

A sentence can express positive, negative or neutral opinion. Here, sentiment score of each sentence is calculated with the help of Vader Sentiment analyzer to distinguish the sentences into positive or negative. Polarity_scores() method is used to obtain sentiment polarity score. Sentiment score of individual sentence will fall in the range of [+1, -1]. Compound score is a metric, which calculates the sum of all lexicon ratings.

Table -1: Compound Score Metric

Sentiment	Scoring
Positive	Compound score>= 0.05
Neutral	(Compound score >= -0.05) and (Compound score < 0.05)
Negative	Compound score <= -0.05

4. IMPLEMENTATION DETAILS

Machine Learning Approach

Machine Learning is a classification method, where the objects are represented by their features. In text analysis, the document is represented based on the words; i.e. words are considered as features. Vectorization will translate the text document into vector features. Different words are assigned with proper feature weights and TF-IDF (Term Frequency – Inverse Document Frequency) is the most common method to allot proper weightage to the words. Test data is converted into TF-IDF feature matrix. The number of feature occurrences will be counted. Words with rare occurrence and words with very high occurrences such as stop words which are very frequent in a text document will be neglected and they will be removed as they does not hold any sentiments. Firstly, feature weights are extracted from the training dataset. Then these features are applied on the test dataset similar to training dataset.

Support Vector Machine: Support Vector Machine is a promising new strategy for the arrangement of both linear and non-linear classification. It is considered as the best classification technique with respect to text classification. It looks for the separating hyperplane. The maximum possible margin between two classes can be considered as best hyperplane. The hyperplane utilizes set of vectors. Support vector machine is a supervised machine learning algorithm which effectively handles text classification problems. Sometimes, it is not possible to linearly classify classes which are closer to one other. Therefore, non-linear classification hyperplane will be used at this point to solve classification problem. SVM is heavily used in analyzing sentiments; Because of the sparse nature of text, it is ideally suitable for SVM.

A. Dataset

For the implementation purpose, data has been scraped from twitter with the help of user credentials. While scraping data, text or tweet is evaluated against Text Blob and Vader sentiment analyzer. As there is no availability of readily available political dataset, there was a necessity to create training dataset manually. But, due to the need of large dataset, it is not possible to manually label text with positive and negative polarity. Therefore, data is scraped and evaluated with natural language processing tools like Text Blob and Vader sentiment analyzer. As per survey[21], Vader sentiment analyzer is more efficient natural language processing tool. Therefore, Vader is considered here to identify text polarity. Then the training dataset is created by analyzing polarity scores and labelled with positive[1] and negative[0] annotations.

B. Data Loading and Model selection

Training dataset is of 2631 tweets. Here, scikit learn tool is used to import the dataset. Real time data is fetched and tested against training data. Learning model is implemented using Support Vector Machine algorithm. SVC( ) function of sklearn python library is used to import the model. Test data size is 800 tweets. Model predicted new test file with an accuracy of 82%. Out of 800 tweets 760 tweets are predicted as negative and 40 tweets are predicted as positive. Text pre-processing is still have to be improved. Graphical representation of prediction gives clear cut idea on this text classification.

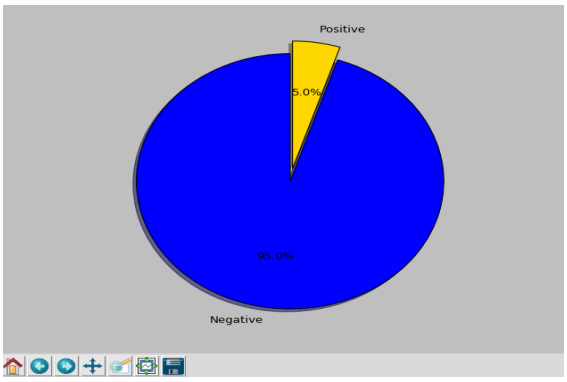

Fig -2: Prediction result

Word Cloud representation of test data is plotted in order to identify the importance of words. This representation helps to identify highlighted words with no proper sentiment values which can be ignored in the pre-processing stage.

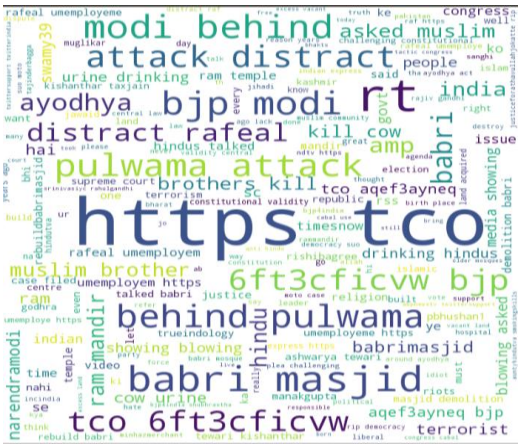

Fig -3: Word Cloud Representation

5. CONCLUSIONS

Proposed system performs an effective sentiment analysis on political reviews of population collected from Twitter. Supervised machine learning technique is used for classification purpose. Sentiment analysis with supervised learning approach, mainly depends on the training set. Ultimately, the problem of Sentiment Analysis is to deal with natural language. It is very difficult to obtain patterns or sentiments from a naturally written language. In the present world, subjective data is continuously growing without limit. Hence, there is a need to keep track of this data which is an extremely good source of information in the process of decision making. Therefore, the proposed system is trying implement a new approach for sentiment classification of text.

In future, the plan is to come up with a technique which enables privacy based sentiment analysis on social media data such as handling political tweets. With the adoption of this algorithm, the system should be able to encrypt the too negative comments, which are harmful to the society in terms of social or political public emotions. Thereby, the proposed system helps in maintaining harmony in the society. Twitter is the largest social media network, which generates millions of tweets daily on variety of issues. Therefore, in the future Hadoop framework which works for distributed Big data, can be used for handling political twitter data.

REFERENCES

[1] Marco Furini, Manuela Montangero, "TSentiment: .On Gamifying Twitter Sentiment Analysis", IEEE ISCC 2016 Workshop: DENVECT, IEEE 2016, ISSN: 978-15090-0679-3/16.

[2] G. Galster, "The mechanism(s) of neighbourhood effects: Theory, evidence and policy implications," in Neighbourhood effects research: New perspectives. 2012, pp. 23–56.

[3] M. Montangero and M. Furini, "Trank: Ranking twitter users according to specific top ics," in Proceedings of the 12th Annual IEEE Consumer Communications and Networking Conference (CCNC 2015), Jan 2015, pp. 767–772.

[4] Walaa Medhat, Ahmed Hassan and Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Shams Engineering Journal (2014) 5, 1093–1113.

[5] B. Liu. Handbook of Natural Language Processing, chapter Sentiment Analysis and Subjectivity. Second edition edition, 2010.

[6] M. Furini, "Users behavior in location-aware services: Digital natives vs digital immigrants," Advances in HumanComputer Interaction, vol. 2014, 2014. [Online]. Available: http://dx.doi.org/10.1155/2011/678165

[7] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth, "The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place," PLoS ONE, vol. 8, no. 5, 2013.

[8] C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, and E. Sulis, "Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in felicitta," in Proc. of the 5th Workshop on Emotion, Social, Signals, Sentiment & Linked Open Data, May 2014.

[9] J. Carrillo de Albornoz, L. Plaza, and P. Gervas, "Sentisense: ´ An easily scalable concept-based affective lexicon for sentiment analysis," in The 8th International Conference on Language Resources and Evaluation (LREC 2012), 2012.

[10] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, "Semeval-2013 task 2: Sentiment analysis in twitter," in Proc. of the Workshop on Semantic Evaluation. Association for Computational Linguistics, June 2013.

[11] Y. H. Hassan Saif, Miriam Fernandez and H. Alani, "Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold," first ESSEM workshop, 2013.

[12] A. khan, B. Baharudin, "Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs," Processed on National Postgraduate Conference (NPC), pp. 1 – 7, 2011.

[13] Judith Sherin Tilsha S., Shobha M S, "A Survey on Twitter Data Analysis Techniques to Extract Public Opinion", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.

[14] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44, Association for Computational Linguistics, 2010.

[15] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.

[16] Upma Kumari, Dinesh Soni, Dr. Arvind K Sharma, "A Cognitive Study of Sentiment Analysis Techniques and Tools: A Survey," in the International Journal of Computer Science and Technology-Volume 8, ISSUE 1, JAN-MARCH 2017.

[17] Seyed-Ali Bahrainian, Andreas Dengel, "Sentiment Analysis using Sentiment Features," in the 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT) - 978-1-4799-2902-3/13 $31.00 © 2013 IEEE.

[18] Devika M D, Sunitha C, Amal Ganesha, "Sentiment Analysis:A Comparative Study On Different Approaches," in the Fourth International Conference on Recent Trends in Computer Science & Engineering. Chennai, Tamil Nadu, India, published by Elsevier B.V.

[19] Deepu S, Pethuru Raj, S.Rajaraajeswari, "A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction", in the 1st International Conference on Innovations in Computing & Networking (ICICN16), CSE, RRCE

[20] Xujuan Zhou, Xiaohui Tao, Jianming Yong, "Sentiment Analysis on Tweets for Social Events", in the Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, 978-1-4673-6085-2/13/$31.00 ©2013 IEEE.

[21] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.