International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

Opinion Mining of Twitter Users using Machine Learning

Khushi Pattanshetty[1], Prashant Tiwary[2], Priyesh[3], Samridhi Shreya[4], Mr. Elaiyaraja P[5]

[5]Department of Computer Science and Engineering, SMVIT, Bangalore

---

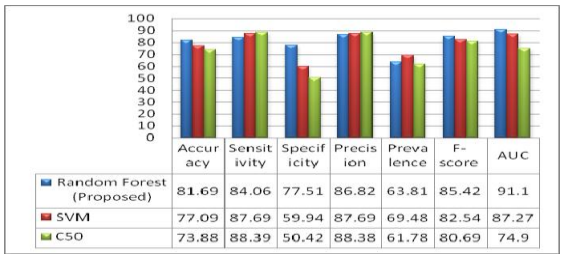
Abstract- Social media services like Facebook, Twitter, LinkedIn, Reddit and many others have become prominent platforms for people across the globe to share their thought, feelings, insight, and emotions in a wide domain spanning across politics, administration, fashion and technology. Amongst these, Twitter is a free and extremely popular public opinion platform and is the best source for collecting textual data. It enables users to post and interact with messages known as tweets. Opinion mining, or sentiment analysis, is a text analysis technique that uses computational linguistics and natural language processing to automatically identify and extract sentiment within text. In this paper we have presented a survey of the user's stance on certain topics in order to understand the concept of opinion mining, real time data collection and learning the different machine learning models.

Keywords- Opinion Mining, Twitter, Pre-processing, Feature Extraction

Introduction- Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing which deals with the detection and classification of sentiments in a text. It is used to sort out people's opening on and their feelings about different subject such as any movie, product, song, etc. and to differentiate between positive, neutral, and negative with respect to different social media platforms such as Facebook, Twitter, etc. Twitter is a free and most popular social networking platform where users can give their opinion in form of tweets. Twitter has changed the current the current system in many dimensions. It has received a lot of attention from researchers in recent times. Around 350 million users post more than 600 million tweets everyday which makes Twitter like a corpus with valuable data for researchers. In this paper we will contribute to the field of sentiment analysis of twitter data using various machine learning algorithms.

Literature Survey-

In 2019, **Arti et al.[17]** dealt with the performance of Twitter's Opinion Mining for Indian Premier League 2016 using the random forest technique shown in fig 1. The data is divided into two parts the training dataset containing 70% and the testing sets containing 30% of the dataset. Tweets are classified into two classes 0 and 1 representing positive and negative opinions respectively. Comparing the actual class and the predicted class values the outcome is decided and depending upon whether true positive, False positive, True negative, False negative values are obtained the classification of the algorithm is checked for its correctness.



	Accuracy	Sensitivity	Specificity	Precision	Prevalence	F-score	AUC
Random Forest (Proposed)	81.69	84.06	77.51	86.82	63.81	85.42	91.1
SVM	77.09	87.69	59.94	87.69	69.48	82.54	87.27
C50	73.88	88.39	50.42	88.38	61.78	80.69	74.9

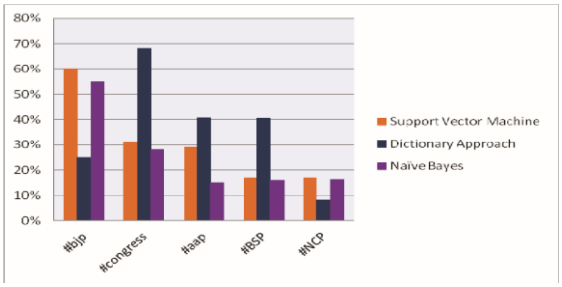
In 2019, **Dhruvi K. Zala et al.[16]** discussed about the N-gram model for prediction of people's opinion on Twitter about current affairs wherein NSR is generated and calculated for N-grams and Hybrid features. An N-gram is a contiguous sequence of n items from a given sample of text or speech. Graphs plotted depicted that the score value was highest for the unigram model followed by unigram model including all features at the feature extraction level, followed by N-gram model and finally it was observed that the N-gram model including all features at the feature extraction level gave the lowest value thereby producing the best result.

In 2018, **Garry Simon et al.[11]** discussed about the influence of Opinion Mining of data collected from Twitter on the market value of a business brand. The methodology applied for opinion mining is Hadoop Map-Reduce Framework, a Map-Reduce algorithm. Map takes a set of data and breaks it into tuples, after that Reduce takes the tuples from map and combines the tuples into a smaller set of tuples. Hadoop Ecosystem consists of two components, namely HDFS (Hadoop Distributed File System) and Map Reduce for Processing. HDFS is designed on a Master-Slave Architecture. The job of a JobTracker is to schedule map and reducing tasks into available slots at one or more TaskTrackers. The client is notified once the conclusion is drawn.

In 2018, **Wiwin Suwarningsih et al.[12]** explained the generation of question-answer pairs from opinion statements collected from twitter in Bahasa Indonesia.

Bahasa Indonesia is a language containing many words from local language. Pang et al. used a supervised learning approach in order to classify data into positive and negative opinions by means of the Naïve Bayes method, Maximum Entropy and SVM. The proposed method covers: pre-processing, sentences extraction, meaning representation, ranking of results, transformation of ranking results and generating QA using the PA template pattern.

In 2018, **Dhruvi K. Zala[10]** dealt with the user's opinion about telecommunication companies with people facing lots of problems in their telecommunication network. Users tweets were retrieved and extraction of emoticons for analysis process was done to increase accuracy. Naïve Bayes algorithm was used for classification of data. Support Vector Machine was used for classification and regression challenges. A decision tree was created for classification or regression models within the type of a tree. K-Nearest Neighbouring was used for info during which the information purposes were divided into totally different categories to predict the classification of a brand-new sample point.



In 2018, **S.Geetha et al.[13]** have proposed Tweet Analysis based on distinct opinions of Social media users. The Future Prediction Architecture Based on Efficient Classification is used which is designed using various algorithms like Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), Artificial Neural Network (ANN) and Fisher's Linear Discriminant Classifier (FLDC). It classifies a tweet as positive, negative or neutral.

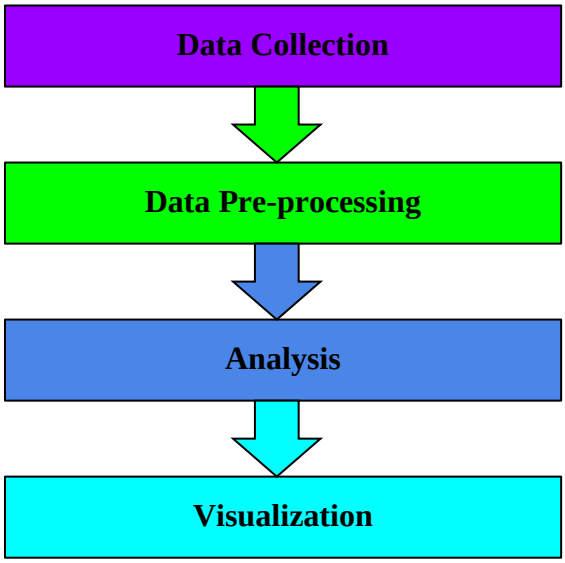
In 2016, **Parul Sharma et al.[6]** predicted the election results from tweets in different Indian languages. Text was pre-processed by removing website URLs, removing hashtags, twitter mentions, emoticons and special characters. The utilization of dictionary based, Naive Bayes and SVM algorithm was done to build classifier and classify the data as positive, negative and neutral. The result of the analysis for Naive Bayes was the BJP and for the Dictionary Approach was the Indian National Congress.

In 2016, **Peter D. Donnelly et al.[4]** have proposed a machine analysis of twitter sentiments to the Sandy Hook shootings. Successfully traced graphs for sentiment analysis which shows a peak of anti-gun feeling on the day of Sandy Hook shooting school which quickly falls to pre-event levels. More surprisingly the analysis shows a peak of pro-gun sentiment on the day of shooting that is sustained at an elevated level for a number of days.

Comparison Table-

AUTHOR NAME	TECHNIQUES	DESCRIPTION
Arti, Kamanksha Prasad Dubey, Sanjy Agrawal (2019)	Random Forest	This method has an accuracy of 81.69% for classification.
Dhruvi K. Zala, Ankita Gandhi (2019)	N-gram model	NSR and Hybrid features gave the maximum accuracy when compared with unigram, bigram and trigram models with hybrid features.
Prof. Pranalini A. Joshi, Garry Simon, Prof. Yogesh P. Murumkar (2018)	Hadoop Map-Reduce Framework	Map breaks data into tuples an Reduce combines tuples into a smaller set of tuples
Wiwin Suwarningsih, Nuryani (2018)	SVM, Maximum Entropy, Multinomial Naïve Bayes, k-Nearest Neighbor.	This method has an accuracy of 81.7% for classification.
Dhruvi K. Zala (2018)	Classification Techniques	Naïve Bayes, Artificial Neural Network, Support Vector Machines, Decision Tree, K-Nearest Neighboring were used to classify the data.
S.Geetha, Vishnu Kumar Kaliappan (2018)	Future Prediction Architecture Based on Efficient Classification (FPAEC)	FPAEC considers complete word context with emotion instead of using particular words and emotion alone providing maximum accuracy.
Parul Sharma, Teng-Sheng Moh (2016)	Machine learning algorithms	Naïve Bayes, SVM and Dictionary approaches were used in which SVM gave highest accuracy of 78.4 and precision of 0.71.
Nan Wang, Peter D. Donnelly (2016)	Machine learning approaches	SVM, Naïve Bayes, Maximum Entropy, Tree, Bagging, Boosting, Random Forest and Neural Network are used and one that gives maximum accuracy for given data set is considered.

Methodology- The step-by-step procedure for analysing public opinion in Twitter is shown below-



Data collection

The process of fetching and measuring information on variables of interest. Data is gathered using Twitter streaming API which will provide twitter feed in a machine readable JSON format.

Data pre-processing

The data collected is messy and full of unnecessary objects which is irrelevant to machine learning classifiers. So, it becomes necessary to first clean the data and remove all redundancies and make it appropriate to feed as input for classifiers. Accuracy of feature extraction also greatly depends on the quality of text data.

Analysis

The main idea behind sentimental analysis is to categorize a tweet as positive or negative. The most important indicators of sentimental analysis are opinion words which imply whether the tweet is positive, negative or neutral.

Visualization

The result from analysis is visualized in the form pie charts or bar charts. A structured analysis helps the users to understand complex data more efficiently.

Conclusion- In this paper, we came across various steps used to perform opinion mining. After evaluating a number of machine learning approaches, we identified those most suitable to classify public opinions. We observed that it is possible to analyze a large set of tweets using various machine learning approaches in a reliable way wherein we employ

A wide range of methodologies to collect, train and classify tweets. After analyzing all the techniques, we wish to deploy the best available algorithms on our dataset and compare the accuracy rates produced and select the most promising one thereafter. In doing so, we would also like to mitigate major challenges faced by sentiment analysis such as sarcasm, negations, word ambiguity, and multipolarity thereby ensuring enhanced accuracy in our future model.

References-

[1] T.K.Das, D.P.Acharjya and M.R.Patra, "Opinion Mining about a Product by Analyzing Public Tweets in Twitter", ICCCI-2014.

[2] Nidhi R. Sharma and Prof. Vidya D. Chitre, "Opinion Mining, Analysis and its Challenges", IJIACS-2014.

[3] Mondher Bouazizi and Tomoaki Ohtsuki, "Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis", IEEE-2015.

[4] Nan Wang and Peter D. Donnelly, "A Machine Learning Analysis of Twitter Sentiment to the Sandy Hook Shootings", IEEE-2016.

[5] Prerna Mishra, Dr. Ranjana Rajnish and Dr.Pankaj Kumar, "Sentiment Analysis of Twitter Data:Case Study on Digital India", InCITe-2016.

[6] Parul Sharma and Teng-Sheng Moh, "Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter", IEEE-2016.

[7] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda and Babita Majhi, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements", SCOPES-2016.

[8] M.Trupthi, Suresh Pabboju and G.Narasimha, "Sentiment Analysis on Twitter using Streaming API", IEEE-2017.

[9] Andrei Pavel, Vasile Palade, Rahat Iqbal and Diana Hintea, "Using short URLs in tweets to improve Twitter opinion mining", IEEE-2017.

[10] Dhruvi K.Zala, "Twitter Based Opinion Mining to perform Analysis on Network Issues of Telecommunication Companies", IEEE-2018.

[11] Prof. Pranalini A. Joshi and Garry Simon, "Generation of Brand/Product Reputation using Twitter Data", ICICET-2018.

[12] Wiwin Suwarningsih, "Opinion QA-Pairs Generation from Indonesian Twitter", IEEE-2018.

[13] S.Geetha and Vishnu Kumar Kaliappan, "Tweet Analysis Based On Distinct Opinion of Social Media Users", ICSNS-2018.

[14] Boppuru Rudra Pratap and K. Ramesha, "Twitter Sentiment for Analysing Different Types of Crimes", IEEE-2018.

[15] Priyansh Sharma, Avruty Agarwal and Neetu Sardana, "Extraction of Influencers Across Twitter Using Credibility and Trend Analysis", IC3-2018.

[16] Dhruvi K. Zala and Ankita Gandhi, "A Twitter Based Opinion Mining to Perform Analysis Geographically", ICOEI-2019.

[17] Arti, Kamanksha Prasad Dubey and Sanjy Agrawal, "An Opinion Mining for Indian Premier League Using Machine Learning Techniques", IEEE-2019.

[18] Andleeb Aslam, Usman Qamar, Reda Ayesha Khan, Pakizah Saqib, Aleena Ahmad and Aiman Qadeer, "Opinion Mining Using Live Twitter Data", IEEE-2019.

[19] Swati Srivastava, Juginder Pal Singh and Deepak Mangal, "Time and Domain Specific Twitter Data Mining for Plastic Ban based on Public Opinion", ICIMIA-2020.

[20] Nithyashree T and Nirmala M.B, "Analysis of the Data from the Twitter account using Machine Learning", IEEE-2020.