iJETRM
International Journal of Engineering Technology Research & Management

SENTIMENT ANALYSIS THROUGH NAÏVE BAYES THEOREM USING MACHINE LEARNING

Rajneesh Sajwan, Yash Kumar Sharma, Vinay Kumar, Amit Sharma
Department of Information Technology, IMS Engineering College, Ghaziabad, India

ABSTRACT
Sentiment Analysis is a way of extraction of people's opinion, intension, sentiments and emotion from his/her text that he has shared on various social media platforms on various topics and social issues. Twitter is one of the such social media platform where people share their opinion in limited text size on various issues that occur in our surroundings. Sentiment Analysis System helps us to classify these opinions according to their sentiments whether they are positive, negative or neutral.
Sentiment analysis has a wide range of applications because opinions are central to almost all human activities and are key influencers of our behaviours. Whenever we need to make a decision, we want to hear others' opinions.

KEYWORDS
Sentiment analysis, Machine Learning, Natural Language Processing (NLP), Naïve Bayes Classifier, Python.

INTRODUCTION

With the emergence of social awareness; popularity and usefulness of social networking such as Twitter increased. Twitter is one of the such important and popular social media platform where anyone can share their views or opinions about any event. In the social networking age people express theirs opining and feelings through Twitter. So, twitter contains huge amount of data. We know that length of any tweets is not more than 140 characters so people can write tweets with correct sentiment/emotions for each word.

Sentiment analysis or opinion mining is nothing but analysis of opinions or emotions from text data. Sentiment analysis identifies opinion or sentiment of each person with respect to specific event. For sentiment analysis we need to pass document or text which can be analysed and generates system or model which represent summarized form of opinion of given document.
Sentiment analysis is a method of analyzing text data to identify its intent.
Sentiment analysis is the process of analyzing online pieces of writing to determine theemotional tone they carry, whether they're positive, negative, or neutral. In simple words, sentiment analysis helps to find the author's attitude towards a topic.
*Positive sentiment may be expressed using words such as "good", "great", "wonderful", and "fantastic".

*Negative sentiment may be expressed using words such as "bad", "terrible", "awful", and "disgusting".

This approach is useful for review of movies, product,
Customer services, opinion about any event etc. This helps us to decide whether the specific product or services are good or bad. Ultimately, this can provide a right direction in order to improve the quality of product/services.

Recent studies on Pulwama attack (which was happened on 14th February 2019) about the sentiments of the netizens of both countries India and Pakistan suggests that both have overall views were negative, but Indian sentiments were more inclined towards hate against Pakistan. These studies were done by using sentiment analysis approach by retrieving the Tweets on this incident.

Following tables shows the sentiments of netizens regarding the Pulwama incident:

iJETRM
International Journal of Engineering Technology Research & Management

Table-1
Overall Sentiments of the netizens regarding Pulwama Attack

	A.Very Negative	B.Moderately Negative	C.Moderately Positive	D.Very Positive
1.#PeaceNotWar OR #Kashmir OR #pakistanleadswithpeace	35.76%	40.14%	19.58%	4.52%
2.#pulawamaTerrorAttack OR #ImranKhan OR #NarendraModi	24.43%	33.45%	29.33%	12.8%
3. #pulwama OR #PulwamaTerror OR #pulwamarevenge	72.18%	25.77%	1.5%	0.55%
4.#PakistanAndCongress OR #PakistanZindabad OR #AirStrike OR #SurgicalStrike	28.14%	13.56%	36.38%	21.92%
5.#pulwama	36.27%	41.65%	12.39%	9.69%
#RemoveArticle370 OR #KashmiriMuslims OR #ExposeDeshdrohis	58.46%	21.54%	11.54%	8.46%

Table-2
Public sentiments from Pakistan

Hashtags	A.Very Negative	B.Moderately Negative	C.Moderately Positive	D.Very Positive
#PeaceNotWar OR #Kashmir OR #PakistanLeadswith Peace OR #Pulwama #PakistanZindabad OR #SurgicalStrikes	29.3%	37.03%	30.43%	3.23%

Table-3
Public sentiments from India

Hashtags	A.Very Negative	B.Moderately Negative	C.Moderately Positive	D.Very Positive
#RemoveArticle370 #ExposeDeshdrohis #PulwamaTerror #pulwamarevenge #surgicalstrike #TerrosristNationPakistan #PKMB	33.09%	40.22%	19.65%	7.04%

iJETRM
International Journal of Engineering Technology Research & Management

Sentiment analysis approaches can be broadly categorized in two classes – lexicon based and machine learning based. Lexicon based approach is unsupervised as it proposes to perform analysis using lexicons and a scoring method to evaluate opinions. Whereas machine learning approach involves use of feature extraction and training the model using feature set and some dataset.

The basic steps for performing sentiment analysis includes:

- **Data collection,**
- **Data processing,**
- **Data Analysis**
- **Data Visualization**

RELATED WORK

Hybrid classification technique has been used for sentiment classification of movies reviews. Integration of different feature sets and classification algorithms such as Naïve Bayes, Genetic algorithm has been carried out to analyse performance on the basis of accuracy. The output of research works shows that hybrid NB-GA is efficient and effective than base classifier and comparing in NB and GA, GA is more efficient than NB. [1]

Polarity of document is also an important aspect in text mining. Future engineering with tree kernel has been discussed by [2]. This technique gives better result than other techniques. In the paper author has define two classification models namely 2-way and 3-way classification. In 2-way classification, sentiments are classified into either positive or negative and in 3-way classification, sentiments are classified into positive, negative or natural. Author considers Tree based representation of tweets in tree kernel method. Tree kernel-based model achieved best accuracy and best feature-based model. Experiment achieves 4% gain than unigram model. [2]

Hierarchical approach for sentiment analysis can be used for cascaded classification [3]. Author cascaded 3 classification-objective versus subjective, polar versus non-polar and positive versus negative to make hierarchical model. This model was compare with 4-way classification (Objective, Neutral, Positive, Negative) model. The output of comparison shows that hierarchical model out perform 4-way classification model. [3]

A domain specific feature based model for movies review has been developed by [4]. Here aspect based technique is used, which analyses text movie reviews and assign sentiment label to it on the basis of aspect. Each aspect is then aggregated from multiple reviews to find sentiment score of specific movie. Author uses SentiWordNet based technique for feature extraction and to compute document level sentiment. The result obtained by algorithm is compared with Alchemy API result. The result of comparison shows feature based model result is better than Alchemy API technique. In short aspect wise sentiment result is better than document wise result. [4]

A huge collection of near about 300000 corpus tweets for sentiment analysis and opinion mining is collected by [5]. A sentiment classifier model is build which identifies tweets positive, negative or neutral. In this technique, collected corpus was divided into 3 sets namely positive emotions- happiness, amusement or joy; Negative emotions- sadness, anger or disappointment and Neutral-text doesn't contains emotions. Tree Tagger is used for POS-tagging for distribution of emotions.

Consumer marketing data is used for collecting sentiments about product and collected data is used for future prediction. Consumer review data is huge amount of data so author uses Hadoop environment for sentiment analysis. Experimental work created Hadoop clusters for analysis of data. Tweets were categorized as positive, negative and neutral [6].

Hadoop's FLUME and HIVE tools are also used for analysis of twitter data. FLUME tool extracts data and stores into HDFS form. HIVE tool is used to extract and analyse data from HDFS type storage. HIVE tool is helps in analysis of different topics by changing keywords. Author identifies sentiments and polarity of tweets from election voting data [7].

Scholars have been conducting a study on sentiment analysis since the last decade which most papers started to appear and rapidly growing after the year 2004 [8]. Sentiment analysis is divided into three different levels which are sentence level,

document level and feature level. The purpose is to classify the opinion either from sentence, document or features into positive and negative sentiment [9]

As for the application of sentiment analysis, it is reported that it has been done in business and marketing, politics and public action context. Example of the application is E-commerce, voting application and world events [10].

SENTIMENT ANALYSIS

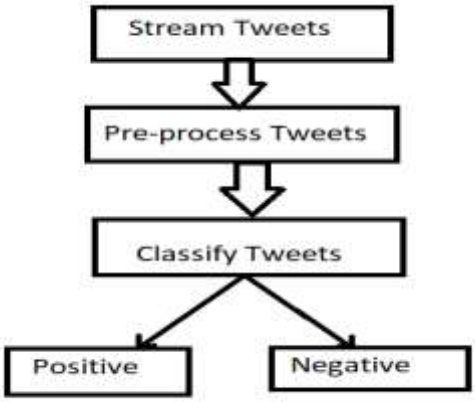
Six different sentiments can be analysed using sentiment package namely anger, disgust, fear, joy, sadness and surprise. By using word cloud frequently occurring words were recorded. A sentiment was added to these frequently occurring words. These new words and sentiments are added to sentiment file for sentiment analysis.
Present uses bayes algorithm. Sentiment analysis algorithm compares each word with words in sentiment file and assigns count for each sentiment Finally it can display count for each sentiment. Present work also finds polarity of text. Polarity will be positive, negative or neutral. In this experiment new words were identified using word cloud and then polarity was assigned to them. Similar to sentiment analysis, it also compares each word with polarity word file and counts polarity of text file. Lastly it displays count for each polarity.

METHODOLOGY

1.Retrieval of Data: Public Twitter data is mined using the existing Twitter APIs for data extraction. Tweets would be selected based on a few chosen keywords pertaining to the domain of our concern, i.e. product reviews. We have elected to use the Twitter API due to ease of data extraction.
2.Pre-processing: In this stage, the data is put through a pre-processing stage inwhich we remove identifying information such as Twitter handles, timestampsof the message and embedded links and videos. Such information is largely irrelevant and may cause false results to be given by our system.

3.Tweet Correction: As tweets are written for human perusal, they often contain slang, misspellings and other irrelevant data. Thus we correct the misspellings in the sentences and look to replace the slang in the sentenceswith words from standard English that may roughly relate to the slang in question. As slang itself can be used to display a wide variety of

sentiment, often with greater emotional impact, this process is necessary so that slang words may be considered as part of the emotion expressed.



iJETRM
International Journal of Engineering Technology Research & Management

NAÏVE BAYES ALGORITHM

Naïve Bayes algorithm is a type of supervised learning algorithm which based on **Bayes theorem**. It is used for solving classification problems. Mainly, it is used in text classification in which a high-dimensional dataset is included.
Naïve Bayes classifier is one of the simple and most effective classification algorithm which helps us to build the fast machine algorithm that can produce quick predictions. Naive Bayes algorithm is basically a probabilistic classifier i.e. it predicts on the basis of the probability of an object.
Some of the main example of the Naïve Bayes Algorithm are: Sentiment Analysis, classifying articles and spam filtration.

EXPERIMENT

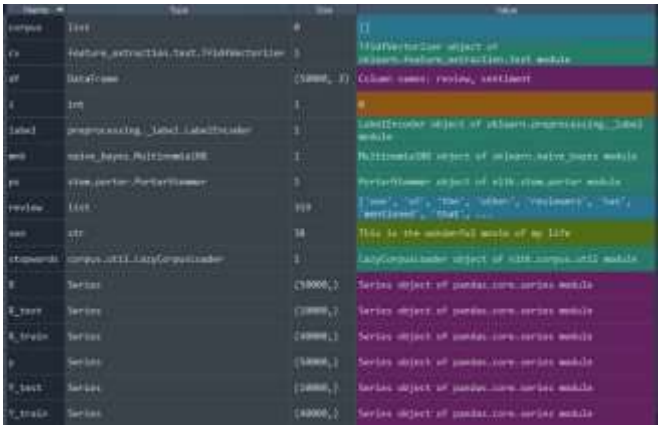
iJETRM
International Journal of Engineering Technology Research & Management

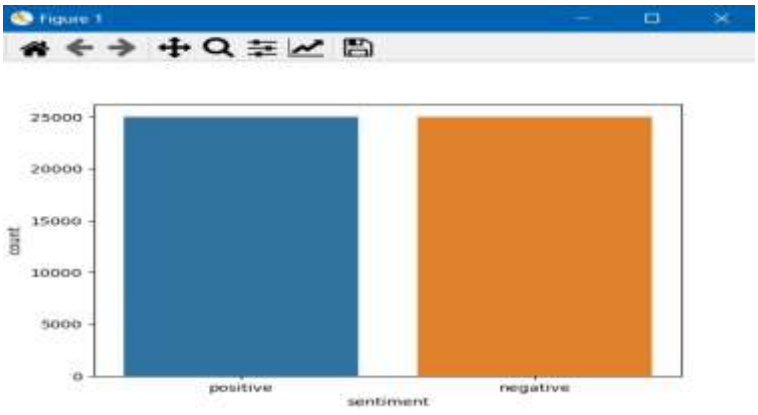
```
model = pickle.load(open('Movies_Review_Classification.pkl','r

def test_model(sentence):
    sen = save_cv.transform([sentence]).toarray()
    res = model.predict(sen)[0]
    if res == 1:
        return 'Positive review'
    else:
        return 'Negative review'

sen = 'This is the wonderful movie of my life'
res = test_model(sen)
print(res)

sen = 'This is the worst movie, I have ever seen in my life'
res = test_model(sen)
print(res)
```





RESULT

As the study goal is to understand the sentimental views of the different comments and reviews which are given on the movies and collected as a dataset of comments. The result of the proposed model of NLP using NLP toolkit, Naïve Bayes and count vector are simulated by anaconda navigator, spyder which uses python 3.6 software with windows 11 (64 bit) O.S. The proposed

ensembles model was implemented in a system containing 8 GB RAM, 1 TB memory with intel Intel(R) UMD graphics 620 and i5 processor 8th gen operating at 1.60 GHz. Results and outputs are obtained after the implementation of the model on the above specification.

REFERENCES

[1] M. Govindarajan, Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume-3 Number-4 Issue-13 December-2013.

[2] Apoorv Agarwal, BoyiXie Ilia Vovsha, Owen Rambow, Rebecca Passonneau, Sentiment Analysis of Twitter Data.

[3] Apoor v Agarwal, Jasneet Singh Sabharwal, End-to-End Sentiment Analysis of Twitter Data, Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, pages 39–44, COLING 2012, Mumbai, December 2012.

[4] V.K. Singh, R. Piryani, A. Uddin, P. Waila, Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification, Conference Paper March 2013, DOI: 10.1109/iMac4s.2013.6526500.

[5] Alexander Pak, Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining

[6] Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari, Sentiment Analysis of Twitter Data Using Hadoop, International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015, ISSN 2091-2730, www.ijergs.org

[7] Sangeeta, Twitter Data Analysis Using FLUME & HIVE on HadoopFrameWork, Special Issue on International Journal of Recent Advances in Engineering & Technology (IJRAET) V-4 I-2 For National Conference on Recent Innovations in Science, Technology & Management (NCRISTM) ISSN (Online): 2347-2812, Gurgaon Institute of Technology and Management, Gurgaon 26th to 27th February 2016.

[8] Mäntylä, Mika V., Daniel Graziotin, and Miikka Kuutila. (2018) "The Evolution of Sentiment Analysis—A Review of Research Topics, Venues, and Top Cited Papers." Computer Science Review 27: 16-32.

[9] N, Mishra, and C. K. Jha. (2012) "Classification of Opinion Mining Techniques." International Journal of Computer Applications 56 (13).

[10] Ebrahimi, M.m Yazdavar, A., and A. Sheth. (2017) "On the Challenges of Sentiment Analysis for Dynamic Events." Intelligent Systems, IEEE 32 (5).