International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

Lexicon Based Sentiment Analysis in Social Media

Shubham Jagdale[1], Mangesh Fule[2], Kedar Jadhav [3], Prathamesh Dhanwade[4], Miss V.B Sutar[5]

B.E. Student, Department of Computer Science and Engineering, Sanjay Ghodawat Institutes, Kolhapur, Maharashtra, India[1]

B.E. Student, Department of Computer Science and Engineering, Sanjay Ghodawat Institutes, Kolhapur, Maharashtra, India[2]

B.E. Student, Department of Computer Science and Engineering, Sanjay Ghodawat Institutes, Kolhapur, Maharashtra, India[3]

B.E. Student, Department of Computer Science and Engineering, Sanjay Ghodawat Institutes, Kolhapur, Maharashtra, India[4]

Assistant Professor, Department of Computer Science and Engineering, Sanjay Ghodawat Institutes, Kolhapur, Maharashtra, India[5]

ABSTRACT: The Information age has created large volume of data as well as opportunities and challenges for users with advancement of internet technology. Most of data present on internet is in unstructured form. This information is created along with social media chats, reviews, posts, comments etc. Platforms like social media -Facebook, Twitter, Instagram, online learning platforms, blogs, gaining popularity as large base of users continuously sharing and exchanging their views, discussions, chat rooms, reviewing products and hence yielding this overwhelming data. Due to this popularity among world we couldn't ignore this mechanism to predict the choices, views and sentiments of large number of users expressing on these platforms. This paper concerned about sentiments expressed through tweets. As Twitter is popular microblogging service with sending limited text. Our aim is to find a way for analysis of twitter data that is helpful for research knowledge for business, endorsements, elections, promotions to extract opinions, sentiments which is extremely unstructured and are either positive or negative or neutral.

KEYWORDS: Natural Language Processing, Opinion mining, Tweets, Machine Learning

I. INTRODUCTION

In recent times, social networking sites used for expressing views, opinions and sentiments. There is lot of data present on internet as large number of peoples have account on social media and through these platforms they share and exchange billions of messages, posts and videos. That's why the main source for generating massive volume of data is social media.While there has been a fair amount of research on how sentiments are expressed in genres such as online reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of microblogging has been much less studied. Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains, but will they also prove useful for sentiment analysis in Twitter? In this paper, we begin to investigate this question.

Another challenge of microblogging sites is the incredible breadth of topic that is covered. It is not an exaggeration to say that people tweet about anything and everything. Therefore, to be able to build systems to mine Twitter sentiment about any given topic, we need a method for quickly identifying data that can be used for training. In this paper, we explore one method for building such data: using Twitter topics (e.g.,'India', 'movies', 'news', 'sports','politics','election','World','life') to identify positive, negative, and neutral tweets to use for training sentiment classifiers. We use various machine learning algorithmstotest sentiment analysis using the extracted features.Depending on individual modelsdid not give a valid accuracy so we pick the top models which gives high accuracy to generate a model ensemble. It is a form of meta learning algorithm method where we assemble different classifiers to improve the predictionaccuracy.Finally,we report our experimental results at the end.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

II. RELATED WORK

Sentiment analysis is research of relations between opinions, expressions to emotion and attitude shows in natural language. Sentiment analysis reached to the peak where it does not only rely on just expressions and opinions a but also deals with behaviour and emotions for different aspects such as language, various topics. In the research of sentiment analysis, researchers used different techniques for predicting the sentiments and emotion through the text such as: Early research work done, on M. Kumar and A. Bala 2015 longer documents, including reviews and blogs. Researchers found that detecting the sentiment in short test is easier than longer document, but difficult to improve the accuracy of sentiment classification in shorter text from microblogging sites. Positive and negative emotions, and hashtags in tweets were studied in as a representative for sentiment label. R. Joshi and R. Tekchandani (2016) Applied the SVM (Support Vector Machine), Naïve Bayesian and the maximum entropy to compare the methods for twitter data analysis, for movie review from twitter for their dataset. R. A. Ramadhani, F. Indriani and D. T. NugrahadiApplied Several methods in Naïve Bayesian smoothing for analysing the sentiment analysis and getting the optimum result 72.3% using the Laplace smoothing technique. Brett Duncan and Yanqing Zhang Applied the neural network to classify the sentiment on tweets. The average accuracy (number of correctly classified tweets divided by the number of incorrectly classified tweets) was 74.15 %.

III. METHODOLOGY

I. **Data Retrieval:** For accessing data from twitter there is twitter API which allows to access its data to twitter developers It is called as Tweepy for simple twitter streaming API. Itallows different ways for fetching tweets through streams: SampleStream and FilterStream. SampleStream provides way to fetch tweets in real time for random short tweets. FilterStream fetch tweets according to the filter i.e. criteria given for accessing data. It can filter the delivered tweets according to three criteria:

- Specific keyword to track/search for in the tweets
- Specific Twitter user according to their name
- Tweets originating from specific location (geo-tagged tweets)

II. **Data Pre-Processing:**Data cleaning is one of the text mining processing to clean the words or other component that is hard to analyse or figure the meaning of the text. words or sentences contains white spaces, punctuations, stop words etc. These characters do not give much information about sentiments and are not needed for the sentiment analysis. Because the data we have collected from the tweets and web blogs are biased and noisy, cleaning data is our first task. Generally, the special characters, such as retweet tags "@RT: xxx" and link address "http://www.", contain less information, they are removed at the beginning also stop words and punctuations are removed by stop word list.

Data pre-processing consists of three steps:

a) Tokenization,
b) Normalization,
c) Part-Of-Speech (POS) tagging.

- **Tokenization**:It is the process of breaking a stream of text up into words, symbols and other meaningful elements called "tokens". Tokens can be separated by whitespace characters and/or punctuation characters.
- **Normalization**: For normalization we identify informal sentences such as all-caps (e.g., *I LOVE this game!!!* and character repetitions (e.g., *I'vegot a mango!! happyyyyyy"*), note their presence in the tweet. All-capitals are made into lower case, and instances of repeated characters are replaced by a single character or removed if it is short word. At last, Twitter tokens is noted (e.g., #hashtags, usertags, and URLs) and placeholders indicating the token type are substituted. Our hope is that this normalization improves the performance of the POS tagger.
- **Part-of-speech tagging:** POS-Tagging is the process of allocating a tag to every word from sentence as to which grammatical part of speech that word belongs to, i.e.verb, adverb, noun, adjective,conjunction etc. For every tweet, we have to count the number of verbs, adverbs, adjectives, nouns, and any other parts of speech.

III. **Sentiment Scoring and Polarity:**Each word gets compared against the lexicon. If the word is found in the lexicon, the sentiment score of that word is added to the total sentiment score of the text.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

For example, the total score of the text: “A masterful[+0.92] film from a master[+1] filmmaker , unique[+1] in its deceptive grimness , compelling[+1] in its fatalist[-0.84] worldview .” will be calculated as follows: ‘total sentiment score’ = +0.92 +1 +1 +1 -0.84 = 3.08, which means, that the text expresses a positive opinion.

The positive polarity of each word was calculated according to the following formula:
the number of occurrences in positive sentences divided by the number of all occurrences.
For example, we calculated that the word “pleasant” appeared 122 times in the positive sentences and 44 times in the negative sentences. According to the formula, the “goodness” score of the word “pleasant” is:

sentScore = #Positive sentences / (#Positive sentences + #Negative sentences)

Similarly, the negative score for the word “pleasant” can be calculated by dividing the number of occurrences in negative sentences by the total number of mentions:

sentScore = #Negative sentences / (#Positive sentences + #Negative sentences)

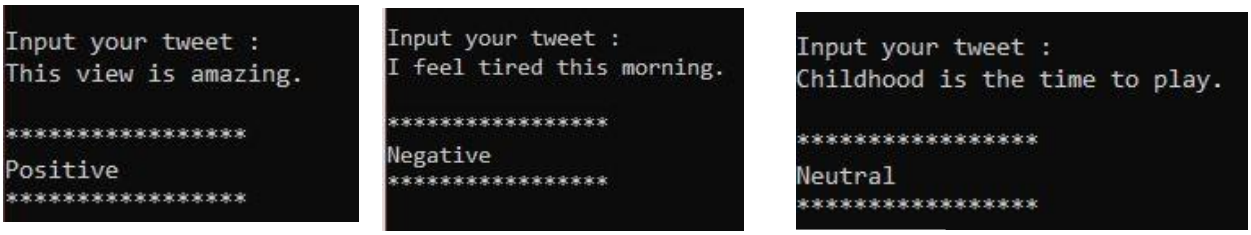
IV. Classification Algorithm:

Algorithms	Accuracy	Completion Speed(s)
Naive Bayes	93.83 %	0.82571
Decision Tree	98.57 %	2.49502
Support Vector Machine	94.25 %	21.54325
K-neighbors	81.03%	5.76202
Random Forest	57.03%	0.80491

The comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and K-neighbors methods for multi-class text classification is presented in this paper.The findings indicate that the Decision tree multi-class classification method hasachievedthehighest(98.57%) classification accuracy in comparison with Naïve Bayes, Random Forest, K-neighbors, and Support Vector Machines classification methods.On the contrary, Random Forest has got the lowest average accuracy values (57.03%) classification accuracy.

IV. EXPERIMENTAL RESULTS

We performed experiments for sentiment classification with classification algorithm as Naïve Bayes, Decision Tree, SVM, K-neighbors and Random forest.This means we performed objective / subjective classification and positive / negative classification. we put criteria while reporting the results of polarity classification (which differentiates between positive, negative and neutral classes) that only subjective labelled tweets are used to calculate these results. However, in case of final classification approach, any such condition is removed and basically both objectivity and polarity classifications are applied to all tweets regardless of whether they are labelled objective or subjective.



(a) (b) (c)

Fig. 2.Sentiment Classification (a) Positive Sentiment (b) Negative Sentiment (c) Neutral Sentiment

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 5, May 2020 ||

V. CONCLUSION

It is an attempt to implement Lexicon based sentiment analysis of data. In this paper we developed a framework for sentiment classification by integration of lexicons and dictionaries. We implemented various classification algorithms and achieved accuracy for our model classification. The system needs to improve the precision in negative cases and recall in neutral cases. In the future, we plan to expand this framework by testing with other datasets.

REFERENCES

[1] R. Joshi and R. Tekchandani, "Comparative analysis of Twitter data using supervised classifiers," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-6. doi: 10.1109/INVENTIVE.2016.7830089

[2] Brett Duncan and Yanqing Zhang, Neural Networks for Sentiment Analysis on Twitter.2017

[3] M. Kumar and A. Bala, "Analyzing Twitter sentiments through big data," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 2628-2631.

[4] R. A. Ramadhani, F. Indriani and D. T. Nugrahadi, "Comparison of Naïve Bayes smoothing methods for Twitter sentiment analysis," 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, 2016, pp. 287-2926.

[5] Alexander Pak and Patrick Paroubek, 2010. Twitter as a corpus for sentiment analysis and opinion mining. In the Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp: 1320-1326b