International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

A STUDY OF SOCIAL SENTIMENT ANALYSIS IN THE TIMES OF COVID -19 USING TWITTER

Princy Sharma[1], Prof. Vibhakar Mansotra[2]

[1]*M.Tech Student, Department of Computer Science and IT, University of Jammu, J&K, India*
[2]*Dean faculty of Mathematical Sciences, Department of Computer Science and IT, University of Jammu, J&K, India*

Abstract: *Sentiment analysis is used to identify user's emotions towards some topics whether it is positive or negative. Social Media is the biggest platform for data sharing where people share their opinion, views and thought related to the current or trending topic. In the current scenario, the fact social media has changed our lives drastically cannot be ignored. Social Sentiment analysis is the use of natural language processing (NLP) to analyse social conversations online and determine deeper context as they apply to a topic, brand, or theme. Twitter is one of the most trending social networking sites and micro blogging site as people express their views on different topics. This paper presents an overview of algorithms of sentiment analysis, techniques, and challenges in the field of social sentiment analysis. Also, a study has been conducted to analyse the sentiments of Twitter users towards sports during lockdown and post lockdown in current times of Covid-19.*

Keywords: Lexicon based method, Machine learning, Natural language processing, Sentiment analysis, Twitter

1. INTRODUCTION

Online social media has created new paradigms of knowledge sharing which not only provides an appropriate platform for the contributors but also active information seekers [1]. Social media sites are providing a facilitative milieu for all to form and broadcast information .The massive growth of technology is pumping a huge amount of data in social media and there are various platforms to share and communicate this data. Micro-blogging has become a standard platform for all online users and has become very frequent in the past few years [2].Within the past epoch, the traffic has almost double on the internet because of various social networking sites [3]. Table 1 shows the statistics of social media's popularity among various age groups [1].

TABLE 1

TOTAL USERS OF SOCIAL NETWORKING SITES	74%
OVERALL TO REQUIRE ADVANTAGE OF MEN	72
OVERALL TO REQUIRE ADVANTAGE OF WOMEN	76

FACT SHEET OF SOCIAL MEDIA USERS

Twitter is a social networking site where people communicate through tweets which means sending short messages to anyone who follows them on Twitter, with the hope that their messages are useful and interesting to someone in their audience. Twitter's strength is real-time and each one the tweets are publicly available and are easily accessible with their geo-tagged locations. Natural language Processing (NLP) is the sub-branch of knowledge science and it's assumed to be vital neighbourhood of data science. It generally teaches machines to read and interpret human-readable texts. It converts information from computer databases or sentiment intents into readable human language. Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Sentiment Analysis of text identifies and extracts subjective information in the source material, and helping a business to know the social sentiment of their brand, product, or service while monitoring online conversations. Sentiment Analysis is that the foremost typical text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative or neutral. With the recent advances in deep learning, the power of algorithms to analyses text has improved considerably. Creative use of advanced AI techniques is often an efficient tool for doing in-depth research[4]. Manual labelling of sentiment words could also be a time-consuming process. Two popular approaches are utilized to automate the tactic of sentiment analysis. The first process makes use of a lexicon of weighted words and thus the second process is based on approaches of machine learning. Lexicon based methods uses a word stock dictionary with opinion words and match a given set of words during a text for locating polarity. As against machine learning methods, this approach doesn't require pre-processed data as it needn't train the classifier.

2. RELATED WORKS

In recent years, different techniques have been applied to perform Sentiment Analysis by various researchers. A review on different techniques in the field of sentiment analysis from past few years is presented as under.

Kouloumpis *et.al.*(2011) performed sentiment analysis using Twitter hashtags (e.g., #bestfeeling, #newphone, #androidwhat) to identify positive, negative, and neutral tweets. Their goal for these experiments was two-fold. First,

they wanted to evaluate whether the training data with labels derived from hashtags and emoticons was useful for training sentiment classifiers for Twitter. Second, they wanted to evaluate the effectiveness of the tweets after pre-processing for sentiment analysis in Twitter data. It was concluded that part-of-speech features may not be useful for sentiment analysis in the microblogging domain [12].

A. Agarwal and JS. Sabharwal (2012) proposed a model for analysis of twitter data where objective was to build a hierarchal cascaded pipeline of models to label a tweet as Neutral, Positive, Negative class using support vector machines with linear classifier and the performance of this hierarchal cascaded pipeline with that of a 4-way classification scheme was compared. Overall, it was concluded that a cascaded design is better than a 4-way classifier design [13].

Stephan Winkler *et.al* (2014) experimented ensemble modelling approach for sentimental analysis using Machine learning algorithms. This approach depend on analysis of words found in sentences and formation of Heterogeneous models (i.e. Binary as well as Multi classification) that were calculated using machine learning techniques i.e decision trees, random forest, neural networks and K-nearest neighbor and are used with boosting algorithm to increase the accuracy of prediction rule. For classifying sentiments based on positive and negative, Gaussian process was used. SVM's was selected the best models from a set of models [14].

Pablo et.al(2014) presented a family of Naive Bayes classifiers for detecting the polarity of English tweets. Two different Naive Bayes classifiers have been built namely Baseline and Binary. The features considered by the classifiers are Lemmas (nouns, verbs, adjectives and adverbs), Multiword, and Polarity Lexicons from different sources and Valence Shifters. The training data set of tweets is obtained from SemEval Organization-2014 and additional annotated tweets from external sources. Many combinations of the above mentioned strategies and features are implemented. It was also concluded that performance was best when binary strategy was used with multiword and valence shifters features [15].

U. Ramakrishnan and R. Shankar(2015) proposed a model based on authorized portals which considered various factors like user behaviour, region, language, popularity, trend and re-tweets. They considered statistical approach to solve the problem of BOW (Bag of Word's) problem. Their methodologies were useful in deriving a resultant which showed the number of positive, negative and neutral tweets generated for a particular topic that has been searched for. The proposed algorithm provided accurate results but takes alot of time for execution [16].

S. Mathapati *et al.*(2016) discussed different approaches of sentiment classification and there performances. They used efficient machine learning algorithms for extracting opinion then they compared the performances of Supervised, semi supervised and unsupervised algorithms with domain adaptation and it was concluded that domain adaptation showed better results than existing learning models by classifying data with different dimensions [17].

Deng *et al.(*2016) proposed a model which used improved recognizing sources of opinions based on new categorization of opinions, i.e., non-participant opinion or participant opinion, in which transudative SVM was used to classify an opinion utilizing existing limited resources. The categorized information was then used by a probabilistic soft logic model to jointly recognize sources of two types of opinion in a single model [18].

S. Shim and M. Pourhomayoun (2017) proposed a method in which data was collected from twitter using Twitters API. Tweets were then converted into java twitter response objects and stored in Mongo Database. They used linear regression model trained by using various supervised and unsupervised machine learning algorithms which showed that the proposed model can successfully predict the gross per day value of the movies [19].

Alsaeedi and M. Z. Khan (2019) provided a relative study for Twitter sentiment which includes machine learning approaches that were ensemble and dictionary (lexicon) based approaches. In addition to that hybrid and ensemble Twitter sentiment analysis techniques were explored. Their research results showed that ensemble Twitter sentiment-analysis methods would perform better than supervised machine learning algorithms. It was also concluded that hybrid methods also performed well and obtained reasonable classification accuracy scores [20].

During the Covid-19(global pandemic) lockdown period, R. Kaur and S. Ranjan(2020) analysed the general public's tweets towards the COVID-19 preventive 21-day lockdown in India from 25th March 2020 to 14th April 2020 and visualized daily sentiment score of Indian user tweets.. The frequency of words and word pairs posted by users in their posts was calculated and it had revealed the overall situation of a region. That can helped policymakers in gauging the situation in the event of a pandemic and planned to cope with it. It was concluded that tacking the respondents can help policymakers in resolving the issues and helping the people [21].

Due to coronavirus pandemic new challenges has been arisen where, K. H. Manguri, et.al (2020) provided a relative study for Twitter sentiments which included machine learning approach (Naïve Bayes). The dataset was related to worldwide COVID-19 outbreaks and they collected data using two keywords coronavirus and covid-19. It was concluded that data regarding the outbreak of coronavirus have showed how people, government organizations and media agencies broadcast the situation [22].

Kavya Suppala and Narasinga Rao in their paper titled "Sentiment Analysis Using Naïve Bayes Classifier" performed sentiment analysis on twitter data by using a Naive Bayesian algorithm. By using model, they could measured the customers opinions and perceptions and enhanced them to any desired level depending on the data gathered from on line resources. In proposed work they used dataset of twitter and facebook[14].

Bac Le and Huy Nguyen in their paper titled "Twitter Sentiment Analysis Using Machine Learning Techniques" built a model to analyze the sentiment on Twitter using machine learning techniques by applying effective feature set and enhances the accuracy i.e., bigram,unigram and object-oriented features. The classification of tweets is done using 2 algorithms i.e., Naïve Bayes classifier and Support vector machines(SVM) whose accuracies are tested by calculating precision, recall and fscore and also shows same accuracy [15].

3. TECHNIQUES USED IN SENTIMENT ANALYSIS

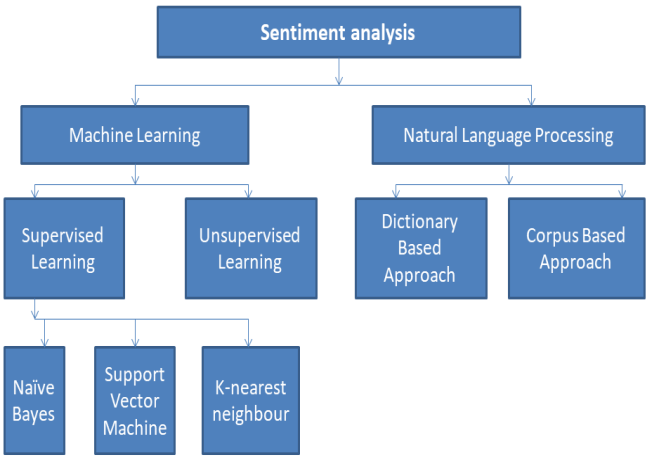


Fig. (1) shows different techniques of sentiment analysis

3.1 Natural language processing

Natural Language processing is a branch of AI and Linguistics, wants to make computers understand the statements or words written in human languages. Natural Language processing came into existence to ease the user's work and to satisfy the wish to speak with the computer in natural languages. Since all the users might not be well-versed in the machine-specific language. It is often defined as a group of rules or a set of symbols. Natural language processing technique plays an important role to urge accurate sentiment analysis. NLP techniques like a part of speech (POS), N-gram algorithms, large sentiment lexicon acquisition, Bag of words, and parsing techniques are used to express an opinion for document level, phrase level, sentences level, and aspect level [9,10]. Large sentiment lexicon acquisition is employed sentiment word dictionary which contains a lot of sentiment words with their numeric threshold value for a particular domain [11]. For instance, consider the subsequent comment. "I just like the bright res". Here "res" refers to resolution, and determination is analogous to graphics. Sometimes textual reviews may contain mixture sentiment

3.2 Machine Learning

Machine learning is a set of statistical techniques and useful for the sentiment classification of text into positive, negative, or neutral categories. Training and testing datasets are required in machine learning. A training dataset is employed to find out the documents and the validation part test dataset is required. There are several machine learning algorithms used to classify text. There are two kinds of machine learning techniques like supervised machine learning algorithms like maximum entropy, SVM, Naïve Bayes, KNN, etc and unsupervised machine learning algorithms like Neural network, Principal Component Analysis, ICA, SVD, etc.

3.3 Naïve Bayes

Naïve Bayes is usually used for document-level classification. The essential idea is to calculate the chances of categories given a test document by using the joint probabilities of words and categories.. Naive Bayes classifiers are computationally fast when making decisions. It doesn't require large amounts of knowledge before learning the Naïve Bayes classifier [12].

3.4 K-Nearest Neighbour

KNN could also be a classifier that relies on the category labels attached to the training documents almost just like the test document. It is how to classify an object that supported the majority class amongst its k-nearest neighbors. KNN algorithm usually uses the Manhattan or the Euclidean Distance. Chebyshev norm or the Mahalanobis distance also can be utilized in this classification.

3.5 Support Vector Machine

SVM may be considered because of the best text classification method. It is a method for statistical classification. Non-linear mapping is performed to maps input feature vectors into a higher-dimensional feature space. It is developed on the principle of structural risk minimization. SVMs can learn a much bigger set of patterns and prepared to scale better, thanks to classification complexity it doesn't depend on the dimensionality of the feature space. SVM has the facility to update the training patterns dynamically whenever there is a replacement pattern during classification [16].

4. APPLICATIONS OF SENTIMENT ANALYSIS

There are various applications of Sentiment analysis. From a user's perspective, people are posting their own content through various social media, such as forums, micro-blogs, or online social networking sites. Sentiment analysis used in

movie reviews, Product reviews, politics, public sentiments and social sites

These applications help us to predict the behaviour of users on the basis of their sentiments shown in the table shown below.

4.1 Movie Reviews

Sentiment analysis has its applications in movie reviews. We can get the information about the movie is good, bad, or average by their star scale rating if a movie is three-star we can predict that movie will be averaged, and if five-star it will be the best review of the movie. Product review: We can predict the identity of different products on the bases of their reviews. People usually want to know other's opinions to check the quality of the product that is good, excellent, average, and poor.

4.2 Product Reviews

Brand monitoring and reputation management is the most common use of sentiment analysis across different markets. Product Review brings additional flexibility and insight into the presentation of the brand and its products. It helps companies to track the perception of the brand by the customers.

4.3 Politics Reviews

Sentiment analysis is becoming an essential part of digital marketing. Monitoring Social media activities have become a prime concern for politicians to understand their social image. Analyses of tweets related to politician leader showcase the opinions of voters before elections.

4.4 Public Sentiments:

Sentiment analysis is useful in the field of gathering data concerning a particular subject which consists of various opinions of the people. People express their emotions in the form of text which is either positive or negative.

4.5 Social Sites

There are so many social networking sites like Facebook, Twitter, Instagram, etc. The users post their views, pictures, and status on these Social Networking sites with the time, the growth of these users goes on increasing. It helps us to predict the personality of the users.

5. CHALLENGES OF SENTIMENT ANALYSIS

Sentiment analysis is a very challenging task. The following are some challenges faced in sentiment analysis are described below:

5.1 Implicit Sentiment and Sarcasm

Most of the studies focus on explicit sentiments and there is a chance that a sentence may contain implicit sentiment even though it is not having any words that earn sentiments. For example, take a sentence having no negative sentiment bearing words but the sentence is negative. Thus, identifying sentiment is important in Sentiment Analysis than syntax detection[23].

5.2 Domain Dependency

In Domain Dependency same sentence or phrase can have different meanings in different domains. For example, the word "unpredictable" is positive in the domain of movies and drama but is negative in the context of the vehicle's steering [24].

5.3 Language Problem

In Emotion AI English language is mostly used because of the availability of its resources means lexicons, dictionaries, and corpora but User get attracted by using Opinion mining with a language other than English like Hindi, French, Chinese, German Arabic, etc. i.e. lexicons dictionaries and corpora for these languages.

5.4 Fake Opinions

Fake opinions or fake reviews misguide the readers by providing them untruthful negative or positive opinions related to any object.

5.5 Negation in Sentiment Analysis

Negation is a Challenging task. Invalidation is communicated from multiple points of view without the reasonable seen utilization of any negative word in the sentence. The presence of negation usually changes the opinion polarity. For example: "It avoids all suspense and predictability found in Hollywood movies." Here the words suspense and predictable bear a negative sentiment, the usage of "avoid" negates their respective sentiments.

6. COVID-19

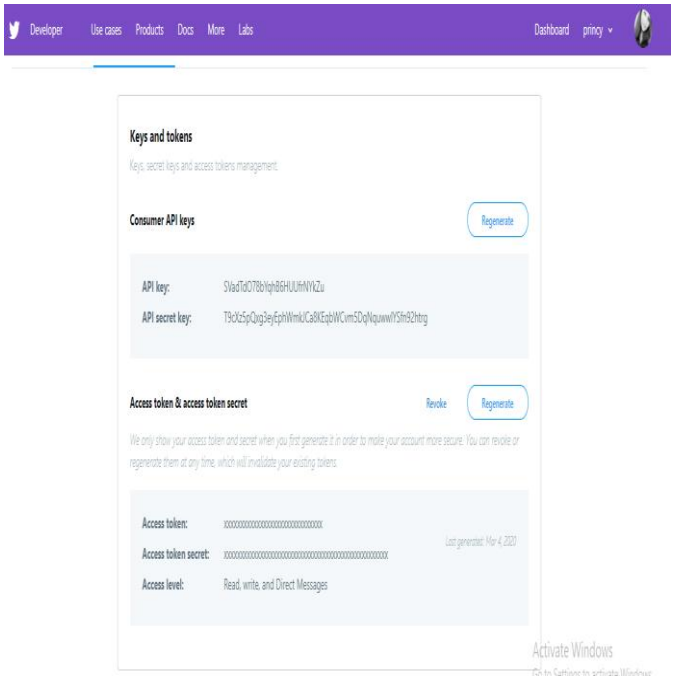
Novel Coronavirus disease or COVID-19 has infected about 4.6 million people worldwide as of mid-May, 2020 . The disease originating from Wuhan, China, has affected most of the countries of the planet .India, like many other countries, has imposed a lockdown on its citizens to slowdown the spread of the disease. The first complete lockdown in India was implemented for 21 days ranging from 24th March 2020. People are restricted to their homes and people who move bent support essential services are observing strict social distancing norms. Public life in India, home to the world's second-largest population during long lockdown periods has witnessed tons of changes. The new social order of self-isolation, quarantine and lockdowns has inspired people to speak via social media platforms. The spread of

stories items and therefore the general public's response to them has been swift because of the social media platforms.

7. A STUDY ON SENTIMENT ANALYSIS OF SPORTS RELATED TWEETS IN THE TIMES OF COVID-19

7.1 Data set description

We started our research by considering general public tweets during lockdown and after lockdown from twitter. In order to collect the tweets a request has been sent electronically through mail to twitter stating the valid reasons for data collection. After the approval of the request, an app has been created on the twitter developer account which has provided keys and tokens to access its data programmatically. We ran the data collection process from 21 March 2020 to 31 March 2020(during lockdown) and from 1 July 2020 to 10 July 2020(post lockdown) which resulted in overall 20,000 tweets in our database. An English language filter was used to collect the tweets. The collected tweets were in JSON format then we converted them into .csv file which contain information such as tweet-text, user id, status source, retweetCount etc. We collected general public related tweets by using R.



Fig(2) keys and tokens provided by twitter

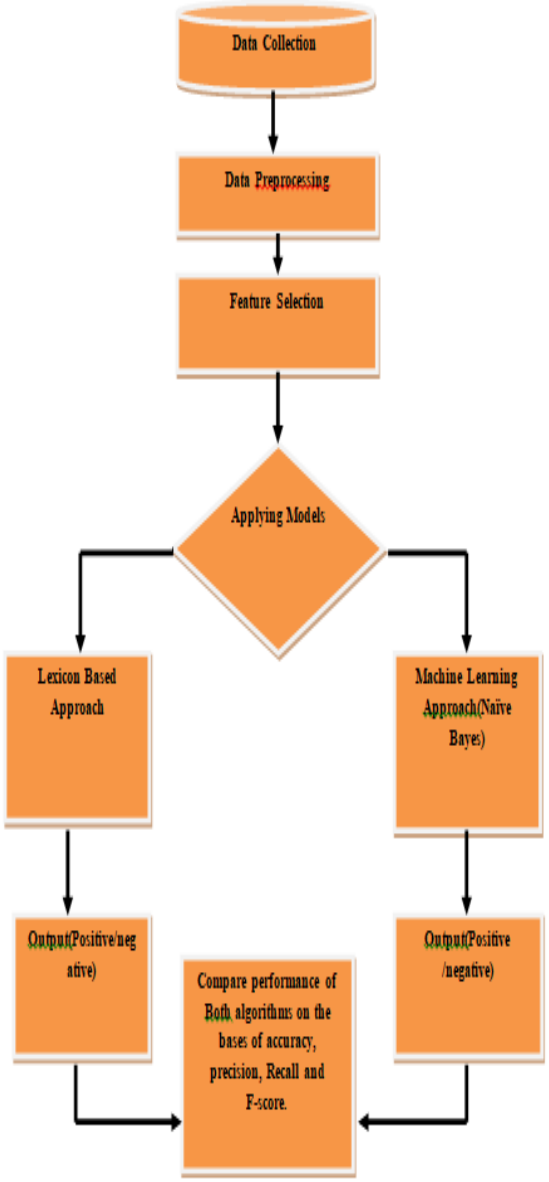


Fig.(3) shows flowchart of proposed methodology

7.2 Data cleaning

The collected tweets were highly susceptible to inconsistency and redundancy. For further processing of tweets, we cleaned the tweets by following various steps: First step was to remove all URLs (e.g. www.xyz.com), hashtags (e.g. #topic), targets (@username) and transform the tweets into lower case. Second step was to remove punctuations, stop words and spelling correction. Third step was tokenization which referred to dividing the text into a sequence of words or sentences. The next task is reducing inflected or derived words through a process called Stemming.

7.3 Lexicon based approach

After data cleaning we used Lexicon Based approach to find out the polarity of tweets either they are positive or negative by subdividing the sentences into words called lexemes. The lexemes were matched with the words of dictionaries by using a rule based technique of POS tagging. Those dictionaries were manually created. The positive word dictionary contains 10,000 words and negative words dictionary contain 8954 words. Thus the polarity of a sentence is the accumulative total (sum) of polarities of the individual words (or phrases) in a sentence. The performance of our approach is shown in Table

7.4 Machine Learning techniques

Machine learning is a set of statistical techniques and useful for the sentiment classification of text into positive and negative categories. Training and testing datasets are required in machine learning. 30 % is for testing and 70% is for training. A training dataset is employed to find out the documents and the validation part test dataset is required. There are several machine learning algorithms used to classify text. These are naïve bayes and random forest and SVM.

7.4.1 Naïve bayes

Naïve Bayes is usually used for document-level classification. In order to apply Naïve Bayes in R .We use a library (e1071) for implentation of naïve bayes in RStudio.

7.4.2 Support Vector Machine

SVM may be considered because of the best text classification method. It is a method for statistical classification. Non-linear mapping is performed to maps input feature vectors into a higher-dimensional feature space. It is developed on the principle of structural risk minimization.

Table 2

Results					
Technique	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Performance**
Lexicon based approach	0.91	0.93	0.92	0.90	0.93
Naïve bayes	0.94	0.96	0.97	0.93	0.94
Support vector machine	0.89	0.90	0.87	0.88	0.86

8. RESULTS AND DISCUSSIONS

The average performance of our proposed is 83.8. After calculating the polarity of during lockdown and post lockdown tweets, we concluded that during lockdown negativity was high due to delay of so many sports events. After lockdown many nations are starting to ease strict lockdown rules, allowing life to return to normal in novel coronavirus pandemic which shows some positivity in sports fans.

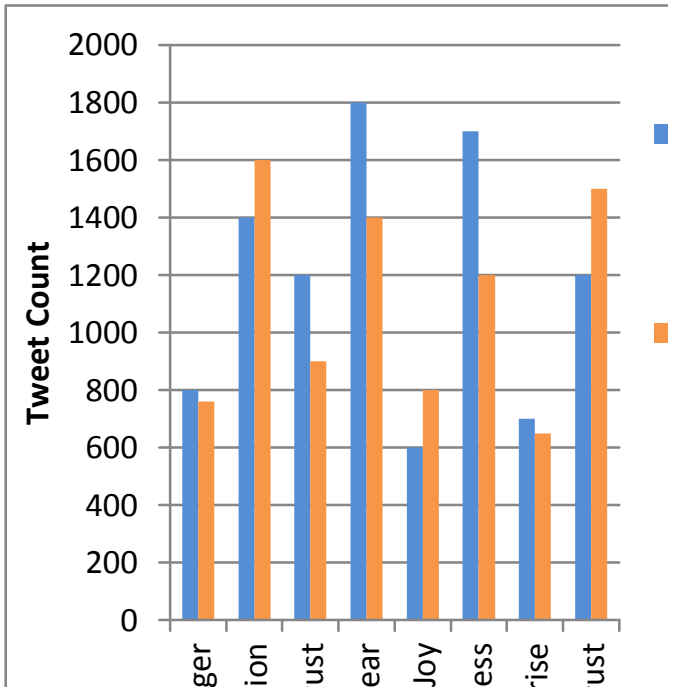


Fig. (4) Tweet count of eight different emotions from during lockdown and after lockdown

Fig. (6) Occurrence of words in lockdown phase

9. CONCLUSION

Sentiment analysis is used to predict human behaviour. A lot of excellent researches has been accomplished in this research area. In this survey, a good deal of state-of-the-art literature has been reviewed.. It also discussed different techniques, application, and challenges of sentiment analysis. The experimentation has been done on Twitter(considering

sports related tweets under COVID-19) by classifying the tweets into positive and negative. Further we calculated precision, recall, f-measure and performance of positive and negative tweets by using different machine learning techniques and lexicon based approach. This study is helpful in analysing the impact of COVID-19 on general public(during lockdown and post lockdown). In the future, more experiments will be performed by using more machine learning techniques.

REFERENCES

[1] T. Singh et.al, "Current Trends in Text Mining for Social Media", International journal of Grid Distributed Computing, Vol. 10, No.6, pp. 11–28, 2017.

[2] L.Williams, et.al, "The role of idioms in sentiment analysis", Expert Systems with Applications, pp. 0957-4174, 2015.

[3] N. Munson, et.al, "Sentiment Analysis on the Social Networks Using Stream Algorithms", Journal of Data Analysis and Information Processing, pp. 60-66, 2014.

[4] K. Shriniwas et al. in their paper titled "Sentiment Classification of News Articles", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5, pp. 4621-4623, 2014.

[5] J. S. Modha, G. S. Pandi, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, pp. 2277 128X, 2013.

[6] R. Varghese, J. M, "A Survey on Sentiment analysis and opinion mining", International Journal of Research in Engineering and Technology(IJRET), Vol. 2, pp. 2319-1163, 2013.

[7] S. B. Moralwar1, S.N. Deshmukh, "Different approaches of sentiment analysis", Computer Sciences and Engineering (ICSE), Vol. 3, pp. 2347-2693, 2015.

[8] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, pp.18-19, 27-28, 44-45, 47, 90-101, 2012.

[9] N. Indurkhya, F. J. Damerau, "Handbook of Natural Language Processing", Second Edition, CRC Press, 2010.

[10] R. Feldman, "Techniques and Application of Sentiment Analysis", Communication of ACM, Vol. 56 No.4, April 2013.

[11] A. Ashari, I. Paryudi and A. M. Tjoa, "Performance Comprison between Naïve Bays, Decision Tree and K-Nearest Neighbour in Searching Alternative Design in an Energy Simulation Tool", International Journal of Advanced Computer Science and Applications(IJACSA), vol. 4, No. 11, 2013.

[12] E. Kouloumpis, T. Wilson and J. Moore, "Twitter Sentiment Analysis: The Good, Bad and the OMG!", 5th International Conference on Weblogs and Social Media, pp. 538-541, 2011.

[13] A. Agarwal and J. S. Sabharwal, "End –To-End Sentiment Analysis of Twitter Data", Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 39-44, 2012.

[14] S. Winkler et al, "Data-based prediction of sentiments using heterogeneous model ensembles", Journal on Soft Computing, pp. 1-12, 2014.

[15] P. Gamallo and M. Garcia, "Citius: A Navie-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 171-175, 2014

[16] U. Ramakrishnan and R. Shankar, "Sentiment Analysis of Twitter Data: Based on User Behavior", International Journal of Applied Engineering Research, Vol. 10, pp. 16291-16301, 2015.

[17] S. Mathapati, S. H. Manjula and Venugopal, "Sentiment Analysis and Opinion Mining From Social Media: A Review", Global Journal of Computer Science and Technology, Vol. 16, pp. 77-90, 2016.

[18] L. Deng, J. Wiebe,"Recognizing opinion sources based on a new categorization of opinion types", International Journal of Computer Science and Information(IJCAI), pp.2775-2781, 2016.

[19] S. Shim and M. Pourhomayoun, "Predicting Movie Market Revenue Using Social Media Data", IEEE International Conference on Information Reuse and Integration, Vol. 04, no.1, 2017.

[20] Alsaeedi and M. Z. khan, "A Study on Sentiment Analysis Techniques Of Twitter Data", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, pp. 361-374, 2019.

[21] R. Kaur, S. Ranjan, "Sentiment Analysis of 21 days COVID-19 Indian lockdown tweets", International Journal of Advance Research in Science and Technology, Vol. 09, issue no.9, 2020.

[22] K. H. Manguri, R. N. Ramadhan and P. R. Mohammed, "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks", Kurdistan Journal of Applied Research (KJAR), ISSN: 2411-7706, pp. 54-63, 2020.