Mamta Gehlot, 2021, 9:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Twitter Data Analysis using Machine Learning Classification Technique

M.Tech. Scholar Mamta Gehlot, Asst. Prof. Vasudha Sharma
Department of Computer Science
LNCT, RIT, Indore. India
gehlotmamta43@gmail.com, vasudhasharma1312@gmail.com

Abstract- Nowadays, maximum people share information through social media around the world. For example, twitter platform where users post their opinions, read posts in form of tweets. such as daily live news, brand, product review, companies review, places and everything because this is a way of interacting with communities. Aim of this paper we are performing a sentiment analysis using collecting real data from twitter API and removing the fake news from kaggle twitter data sets. which are more difficult to analyze? Data will be preprocessed with tokenizations and stop words will be removed from twitter data set of given input data set. After then applied feature extraction algorithms of given input data and priority to each word is assigned to classify positive, negative. After that the noisy data into several models for training of data. For this sentiment analysis we are comparing different machine learning classifiers with twitter data in this paper. In the KFC and McDonal's both are data sets which are more popular subjects and 14000 tweets for our research. we have used 10000 tweets for training purposes and 4000 tweets for testing purposes. Find the result from these models was tested using different parameters from our proposed method. Using this study, we can contribute to the field of sentiment analysis by analyzing performance.

Keywords:- Sentiment Analysis, Twitter Data, Social Media, Machine Learning Classifier: Support Vector Machine (SVM), Decision Tree (DTs), RandomForest (RF).

I. INTRODUCTION

In recent year people in these days depending on microblogging sites like facebook instagram, tumblr, twitter youtube milions people share the posts, live news and express their opinion about different subjects such as a political affair, product review, educational, women issue and general topics, extracting knowledge from the twitter data [1]. Sentiment analysis is the mining of opinion and analysis of twitter data and that describe as a positive negative and neutral category which explore data from various social media platforms [2]. The aim of this analysis in research determining the subjectivity opinion. Result of this analysis based on This sentiment analysis and review of tweets or classified opinions which are based on the data size and document type [4]. Twitter application is an excellent medium for creation of tweets presentations [5].

Twitter analysis is a popular topic for research. Such analysis is useful because it's gathering by crawler data which are used for collect to data from twitter and classified public opinion by analyzing of vast social media data [6]. The aims of this study that analyse the level of sentiment from the social media [7] In this sentiment analysis we are using twitter API for extracting data then cleaning the data and after these processes fed data into three classified tweets on the basis of sentiment (new data) [8]. This Analysis

International Journal of Science, Engineering and Technology

An Open Access Journal

helps to understand the way of thinking about any research topic brands, products etc [9].

Through the advertisement campaign can see how people are reacting from this campaign in personal marketing. There is a way to analyze sentiment related to them [10]. Use of the same campaign can be seen as reacting for Political parties and can be analyzed.

There are several reasons for sentiment analysis where we can choose twitter data as given below.

- On twitter more than 500 million number of tweets on daily bases and that is a vast level of data for sentiment analysis.
- On Twitter there are number of all age groups people, with a high percentage of business executives' people being present from many countries on social media.
- 50 million or more people download from many browser twitter applications.

In this study we have used of supervised learning Classifiers to analyze the sentiment of the people for this analysis. Such as Support Vector machine learning classifiers (SVM), Decision tree (DTs) and Random forest (RF). In this result we will compare all classifiers based on accuracy which gives the best result. Finally for this research we also used machine learning techniques.

In our work we introduced of score vector of tweets and our external features with n-gram of features and show that impact of SVM classfires on for improve our classification performance level.

This study defined the concept of opinion based expression in the sentiment analysis of twitter in the field of machine learning. The model which is proposed using multiple algorithm to enhance the accuracy of the classification of tweets. The analysis of twitter data is being done in various aspects to mine the sentiment.

Sentiment analysis deals with opinion classified into positive, negative, neutral. The proposed model involves a supervised and unsupervised algorithm. After fed data into a supervised model for testing and classification of an entity with the highest accuracy.Show that Result of this analysis using machine learning classifier such as SVM has the highest accuracy and random forest, Decision tree is very effective.

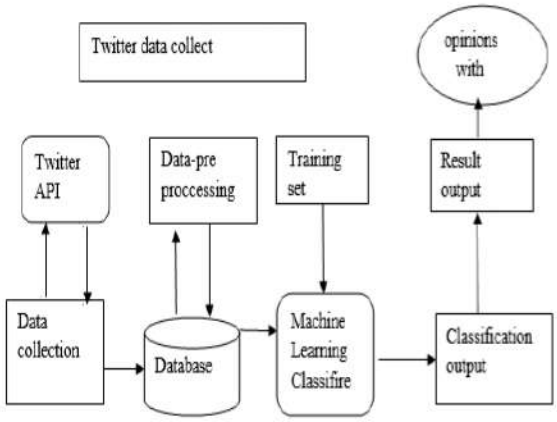

Fig 1. Working Architecture of Twitter classification.

II. LITERETURE WORK

Santhosh Kumar et all., IEEE, 2016. The study of mining and Analysis of the opinion on the twitter data can explore things from various social platforms using machine learning techniques. Through the twitter platforms user sent to- post - read post this is known as a ' tweets'. theirs is the way of share the information and opinion and that's comments on other post.

This is the good platforms for advertisement of other tweets and opinion. In this model extract the twitter data from the social media and removed the unwanted data and classified into three categories negative, positive, neutral. For this paper we will collect hotel data from the twitter for this analysis using machine learning algorithms and find the result from the various metrics for the get accurate output result.opinion based analysis using different classifier for to get accurate result of tweets. In this model both supervised and unsupervised machine learning algorithms [9].

L. L. Bo Pang et all In this paper we have done used of different deep learning methods for this sentiment analysis with twitter data. Through the deep learning technique can at the same time solve a wide range of problems or complex operations and gain popularity among researchers. Deep learning algorithms by themselves generate the high order features to predict respect of the object in this feature extraction. Which is helpful in generating respect for

the object. Using Features of deep learning can handle huge amount of structure and unstructured data. Generally utilize of two types of neural network, Convolution neural network (CNN) for the use of image processing and recurrent neural network (RNN) with nature language processing. We can use of different type of embedded system such as a word2Vec, global vector (GloVe). Various combinations are applied for the best score value for each model and compare themperformance [10].

V. Lakshmi et all in this analysis we collect data with opinion in this new era from microblogging sites such as twitter, facebook, Twitter is a platform where people share their information, ideas as tweets and twitter is one of the good sources for sentiment Analysis. People Opinion can be divided into three categories: negative, positive, neutral and performing analyzing different types of opinion, grouping that is ne of ccessary for sentiment analysis. Data mining are used for unwanted information from social networking sites and text mining, natural language processing are used for it.

The aim of this research classification of the tweets using the machine learning techniques to improve classification results for sentiment analysis and increasing efficiency and reliability of propose approaches. Using decision tree, hybride, Ad boosted tree to gets highest accuracy of the classifier.This proposed model based on the two preprocessing stage and classifier stage. Hybrid Models are used to improve classifier accuracy and f-Measured[11].

Levy, M et all Today's people share the information through social media site like facebook, twitter in the world. Twitter is a platform which is used to interact with different communities.where users send posts-read posts known as tweets.user updated opinion such as daily news ,brand, various places. Aim of this model collocet the real data from twitter account and then performed sentiment analysis. Our work for this model we are using both supervised and unsupervised machine learning algorithms.For sentiment analysis we are performing using extract the data directly from twitter API, then cleaning process and discovery of the data. After fed data will be into a several models for the purpose of training . Each tweet classified into three catogories negative, positive and neutral on the basis of sentiment analysis. Where data collect two subject MCdonald and KFC.

Which are more famous. For these models use different machine learning algorithms and find the results using cross validation, f-score, maximum entropy, various testing metrics etc. In further work use of the same methodology for various fields like detecting, rumors on twitter regarding the spread of diseases [12].

W. Medhat, A hussan et all Today's the modern era based on the internet where people share the opinion, idea;s through social media such as : microblogging sites, personal blog, reviews and so on. One of this sentiment analysis is a part of text mining were Analyzed of people opinion and divided into tweets as good, bad, neutral. In this paper work published of people reviews and sentiment of tweets then identified by searching using a particular keyword and evaluate the polarity based of tweets as a positive negative. Using a Naive Bayes classifiers (NBC) can be both test the data and features of words or also evaluate the Polarity of sentiment of each tweets. Compare three machine learning classifiers, namely, Random forest, Naive Bayes and support vector machine with performance evaluate parameters such as accuracy, precision. Using Classifier such as RF, NBC, SVM increasing both estimated accuracy and three features of the number of tweets. In future work some more features can be added which are used for improving accuracy of prediction[13].

Sidharth,darsini et all Everyday, globally people are share their ideas and information using social media platforms.Twitter application is one of the most popular platforms for sharing of the opinion, reviews, posts and particular topic issues. The main focus of this paper that performed sentiment analysis of people opinion and social issues of women's by this proposed model which is very critical problem In many countries with every woman. Using twitter scraper collecting data from twitter to build a dataset in python programming then clean the data set and remove noise from the data set. python programing tools like text blob which are used for classified of each tweets as and technology are used.By theText blob to classified each tweets as a good,bad and neutral based on polarity of sentiment. #Women and #MeToo. There are two data sets hashtag. Through the different machine learning algorithms can be tested on the model. After Results compare the performance of each model with tested data using various testing parameters like precision, recall, f1-

score. Support vector machine are used for higher accuracy of both #hashtag (#Women's and #Metoo). #women have more popular hashtag then #Metoo shares information. In future work the same methodology can be used in various fields like product review in sentiment analysis [14].

Harpreet Kaur et all in this analysis is to predict the polarity of the word and classify them according to negative, positive tweets. For this paper two types of classifiers are used namely lexicon based and machine learning. Multinomial naive bayes (MNB), Support vector machine (SVM), Logistic Regression (LR)Recurrent Neural network (RNN).

In this paper existing data sets have been used first one "Sentiment140" from stanford university, which are consisting of 1.6 million tweets and other one original consisting of every library data 13870 entries by 'Crowdflower' data and both data sets are classified based on sentiment. Various classifiers are performed on both data sets and obtained as a result then compared with them. Use of this model for sentiment to predict a new data. By using a Text machine learning models data will be trained and accurately classified based on standard dictionaries [15].

B. O'Connor et all in recent years twitter is a more popular topic for sentiment analysis. In this analysis of tweets blessed on ordinal regression. Aim of this analysis perform the sentiment analysis of tweets which are based on ordinal regression with machine learning techniques. In this approach using of feature extraction method on pre-processing tweets and creates an efficient feature. For feature scoring and balancing can be used in several classes. Supervised learning classifier such as Multinomial logistics regression (MLR), Support vector machines (SVM), Decision tree (DsT), Random forest (RF) are used in this research. In this analysis using NLTK corpora resources twitter data set publicly made available for implementation of this system. Using machine learning methods.

Finding that detection of ordinal regression with best accuracy for Experimental. However obtain best performance using Decision tree over the other methods. For result Decision tree give the high accuracy at 91.81% with Mean Absolute Error and Mean Square error. For the future work use of bigramS and trigram with different deep learning techniques are used in improve approaches [16].

III. PROPOSED METHODOLOGY

In this study, a number of datasets have been applied for training purposes as well as testing using the support vector machine learning algorithm and then compute the polarity of each sentiment or reviews.The sentiment analysis techniques contained various steps and these steps are:

1. Input data:
Input data is the twitter data in the first steps which is given as a real time data using a twitty application which is extracted through the various social networking sites and microblogging websites.

2. Pre-processing:
In the preprocessing phase twitter stream will extract all twitter data which is an unstructured form of data that is given as input. tweets will be preprocessed with tokenizations and stop words will be removed from data.

3. Feature Extraction:
Third step of feature extraction after use of this algorithm the pre-processed data will be given as input where n- gram algorithm has been applied and then assigned to priority of each word which is needed to classify.

4. Classification:
In this paper classification will be using three classification algorithms/techniques for this sentiment analysis. Classification can be applied after application of feature extraction algorithm for this sentiment analysis. work of this study, SVM, random forest, decision tree can be applied for this sentiment analysis.

4.1 Pseudo code of N-gram algorithm:

- **Input:** Tokenlized strings TS, Matched Strings MS
- **Output:** Similarity list (CS)
- Construct dictionary of N-gram based on TS
- Traverse the input query string S into the candidate the N -gram list TS
- Set the MS matched strings = 0;
- For each input string belongs to Ts
- Find the input string from each words TS
- For each input string belongs to TS
- Frequency = frequency + 1;
- If (frequency>threshold)

- Put the input string in candidate list (CL)
- For each Z belong to the candidate list CL do.
- Calculate similarity (input string Z).
- **Results:** calculate similarity (CS).

4.2 Pseudo code of SVM classifier:

- **Input:** Calculated similarity list (CS)
- **Output:** Classified data
- Weight=0 bias=0 input=0
- R=max(x)
- While the whole data get classified into two classes in the for loop do
- For i=1 to Cs(n) do
- If Yi(<WiXi>+bias) <0 then
- Wk+1 = Wk+YiXi
- K=K+1;
- End if
- End while
- Return Classified data K, is the number of the classes and x is the data in the classes

4.3 Pseudo code of Random Forest classifier:

- **Input:** Dataset (tweets) for training and testing
- **Output:** Classification result interns of Accuracy, Precision, Recall, and F-measure
- begin
- Preprocessing and normalization of data;
- for Training data set to do
- Calculation of features at the start of the given training set select any K random data points.

  For these K values find the decision trees for the dataset given. After that we select the N number of trees that we want to create. Now for the new data points obtain the predictions from the decision trees created, after this, we assign the particular points of data that are selected earlier to every tree that is getting the maximum votes.
- Build classifiers;
- end
- Use the value of features for respective tweet;
- for all records in testing data set to do
- Check accuracy of the model;
- end# Training and Testing
- end

4.4 Pseudo code of Decision Tree classifier:

- **Input:** Dataset (tweets) for training and testing
- **Output:** Classified result intern of Accuracy, Precision, Recall, and F-measure.
- begin
- Preprocessing and normalization of data;
- for Training data set to do
- Calculation of features
- Apply Decision Tree training;
- Build classifiers;
- end
- Use the value of features for respective tweet;
- for all records in testing data set to do
- Check accuracy of the model;
- end# Training and Testing
- end

4.5 Flow Chart of Proposed Work:

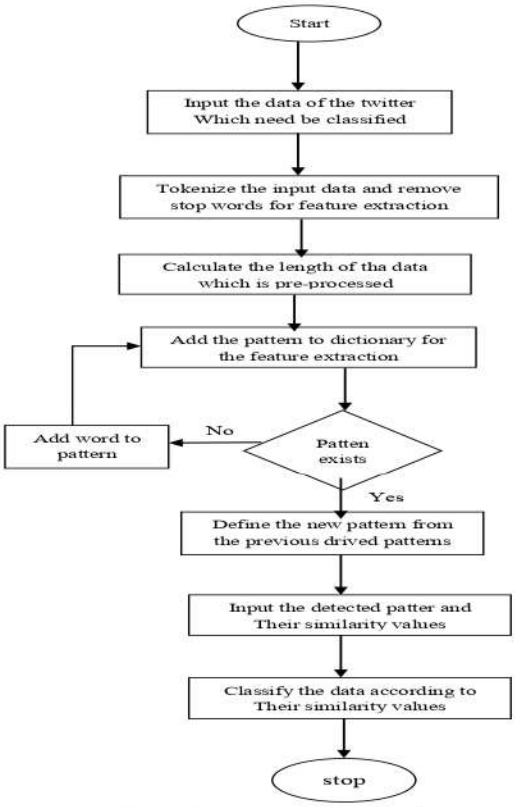

Fig 2. Proposed Flowchart.

IV. IMPLEMENTATION

1. Data Set Processing:

This particular section provides the details of the experiments that we have performed for the analysis of the proposed methodology in the context of twitter analysis. We have done tests on the Kaggle twitter data set. Data set based on the challenge launched by the KFC, and McDonald's of AI - Algiers, which consists of building a system that can classify tweets as Sad or Happy. Currently, we have check tweets that are correct or incorrect. In the KFC, and McDonald's dataset, we have taken 14000 tweets for our research. 10000 tweets for training and 4000

International Journal of Science, Engineering and Technology

An Open Access Journal

tweets for testing purposes. Data set link mention below.https://www.kaggle.com/mcdonalds/nutrition-facts, details figure 3 and figure 4.

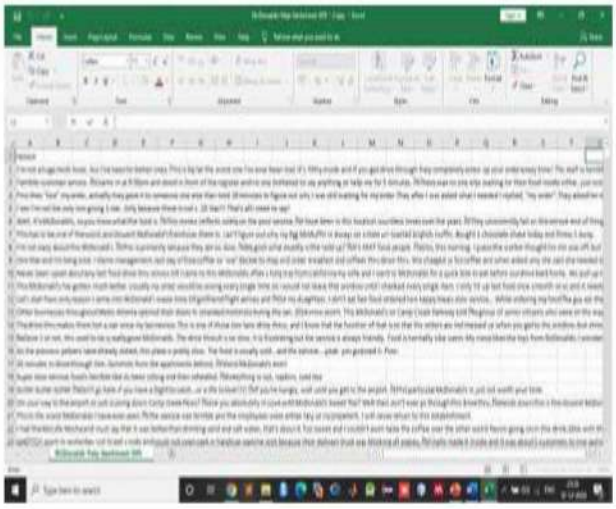

Fig 3. Without Processed Data.

There are URL's, own usernames, special characters and repeated words and symbols. Then we have to remove all the Hashtags identified by the # symbols, all the special characters, URL's, own usernames and repeated words.

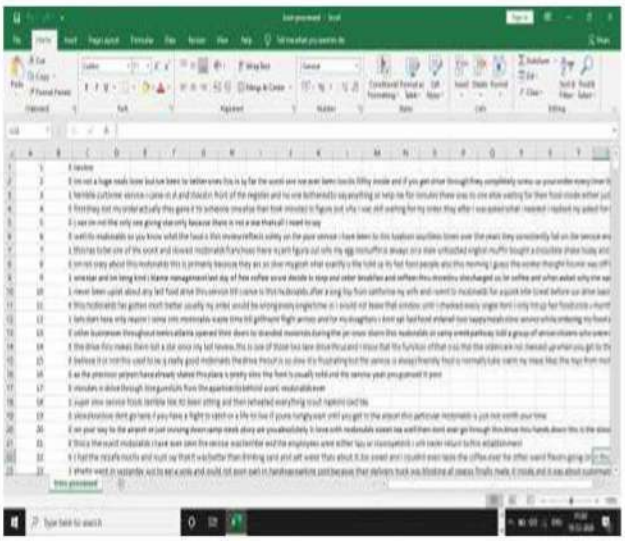

Fig 4. After Processed Data.

V. RESULT

Our proposed method evaluated in below parameters:

- Recall
- Precision
- Accuracy
- F1-Score

Recall= tp/(tp+fn)..... (1)

Precision= tp/(tp+fp) .....(2)

Accuracy= (tp+tn)/(tp+tn+fp+fn)..... (3)

Table 1. Evaluate Metric with Contingency Table.

		Prediction	
		Predicted Negative	Predicted Positive
Reality	Actually Negative	True Negative (TN)	False Positive (FP)
	Actually Positive	False Negative (FN)	True Positive (TP)

1. Recall:

Recall evaluates the quantity of positive class expectations made out of every single positive model in the dataset. The recall is calculated using equation 1.

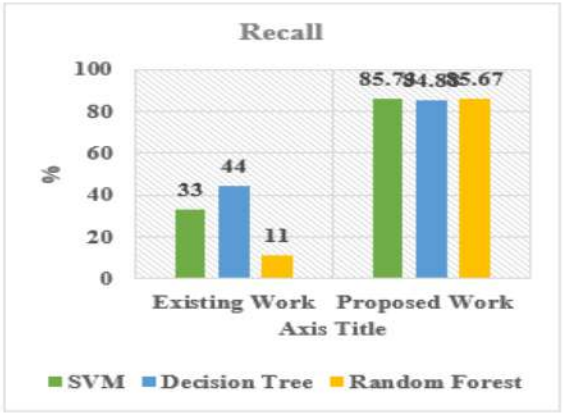

Fig 5. Recall Figure between Existing Work and Proposed Work.

In Figure 5 we calculate recall value for all algorithms Existing Work and Proposed Work. And results shown with the help of a diagram. We find that the proposed approach shows truer positive rate compare to collaborative and content-based approaches.

2. Precision:

Precision measures the quantity of positive class expectations that really have a place with the positive class. The precision is calculated using equation 2.

International Journal of Science, Engineering and Technology

An Open Access Journal

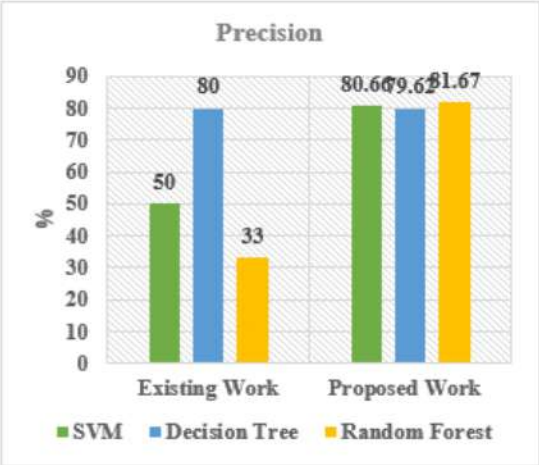

Fig 6. Precision Figures between Existing Work and Proposed Work.

In Figure 6 we calculate a precision value for all algorithms Existing Work and Proposed Work. And results shown with the help of a diagram. We find that the proposed approach shows truer positive rate compare to collaborative and content-based approaches.

3. Accuracy:

Accuracy is essentially a proportion of the accurately anticipated groupings (both True Positives + True Negatives) to the absolute Test Dataset. The accuracy is calculated using equation 3.

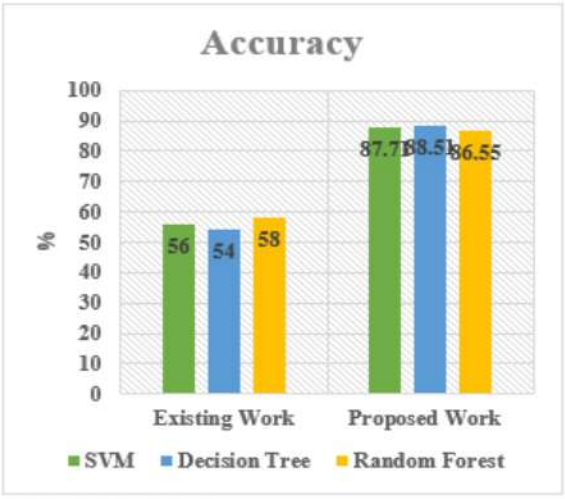

Fig 7. Accuracy Figures between Existing Work and Proposed Work.

In Figure 7 we calculate a precision value for all algorithms Existing Work and Proposed Work. And results shown with the help of a diagram. We find that the proposed approach shows truer positive rate compare to collaborative and content-based approaches.

4. F1_Measure:

The accuracy of the test is stated by the F1 measure, this is specially used in the binary classification. The precision and the recall are used for the calculation of the F1 measure.

All samples that should have been identified as positive on Figure 8.

F1_Score = 2 * ((precision * Recall) / (precision * Recall))

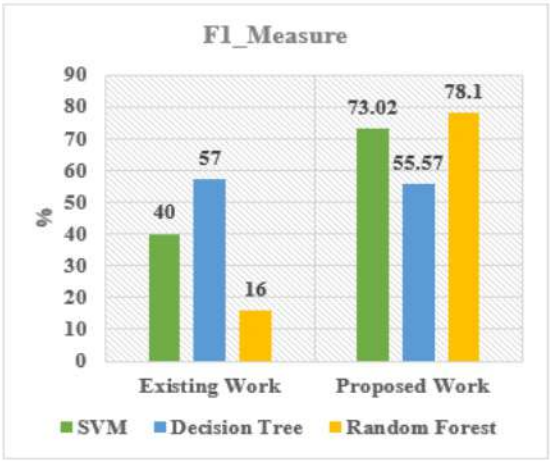

Fig 8. F1_Score Figures between Existing Work and Proposed Work.

Figure 8 here we can say from the above results that the proposed approach is efficient. And running time is reduced to an extent by keeping the quality of recommendation as to its best. This concludes that the proposed method is scalable and can be applied to a large dataset.

VI. CONCLUSION

In this paper analysis of different opinion based expression using machine learning techniques based such as Support vector machines (SVM), Random Forest (RF), Decision Tree (DTs) and so on in the social media.

Through several methods to enhance accuracy of classification of our proposed models into tweets positive, negative, neutral. After fed data into supervised model for training and testing of new data sets that using this model we get a highest accuracy of classifires.

In our proposed model included supervised and unsupervised both algorithms. We have also used of various testing parameter or machine learning classification algorithams have used for tested and trained data sets or evaluate performance of the classifires.such as SVM , DTs, RF.

Show that in result SVM has a greatest accuracy for both subjects KFC and McDonal's which are more popular in the term of Recall, F1-score.

In future work use the same methodology in various fields like detecting rumors and use of different tweets with different data sets. We get the highest accuracy with highest performance for the best result.

REFERENCE

[1] El Rahman, S. A., AlOtaibi, F. A., & AlShehri, W. A. (2019). Sentiment Analysis of Twitter Data. 2019 International Conference on Computer and Information Sciences (ICCIS).

[2] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi and Wejdan Abdullah AlShehri, Sentiment Analysis of Twitter Data, 978-1-5386-8125-1/19/ $31.00 ©2019 IEEE

[3] Rekha V, Raksha R, Pradnya Patil, Swaras N and Rajat GL, Sentiment Analysis on Indian Government Schemes Using Twitter data, 978-1-5386-9319-3/19/$31.00 ©2019 IEEE

[4] Sonia Saini, Ritu Punhani, Ruchika Bathla and Vinod Kumar Shukla, 2019 International Conference on Automation, Computational and Technology Management (ICACTM) Amity University, Sentiment Analysis on Twitter Data using R.

[5] Nann Hwan Khun and Hninn Aye Thant, Visualization of Twitter Sentiment during the Period of US Banned Huawei.

[6] Sani Kamış and Dionysis Goularas, 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data

[7] Lei Wang, Jianwei Niu, and Shui Yu, SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis.. Journal of Latex Class Files, Vol. 14, No. 8, August 2018.

[8] Alaa S. Al Shammari Real-time Twitter Sentiment Analysis using 3-way classifier, 978-1-5386-4110-1/18/$31.00 ©2018 IEEE.

[9] Santhosh Kumar K L and Jayanti Desai, Jharna Majumdar, "Opinion Mining and Sentiment Analysis on Online Customer Review" In 2016 IEEE International Intelligence and computer Research.

[10] L. L. Bo Pang, «Opinion Mining and Sentiment Analysis Bo», Found. Trends® Inf. Retr., vol. 1, no. 2, pp. 91–231, 2008.

[11] V. Lakshmi, K. Harika, H. Bavishya, Ch. Sri Harsha, "Sentiment Analysis of Twitter Data," vol. 04, February 2017. Link- "https://www.irjet.net/archives/V4/i3/IRJET-V4I3581.pdf"

[12] Levy, M (2016). Playing with twitter data. [Blog] R-bloggers. Available at: http://www.r-bloggers.com/playing-with-twitter data/[Access].

[13] W. Medhat, A hussan, and H. Korashy. "Sentiment analysis algoritham and application: A survey," Ain shams Engineering gournal, vol. 5 no.4, app1093-1113,2014.

[14] Sidharth, darsini, and sujithra," Sentiment Analysis on youtube & Twitter Data Using Machine." Int. j. Res. Appl. Sci. Eng.Technol., vol. 8 no.5pp. 755-758,2020.

[15] Harpreet Kaur, Veenu Mangat, Nidhi, "A survey of sentiment analysis techniques", February 2017,2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

[16] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.," Icwsm, vol. 11, no. 122–129, pp. 1–2, 2010.