International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 1, Jan-Feb-2021, ISSN (Online): 2395-566X

# Sentiment Analysis of Twitter Data by Making use of Machine Learning Algorithm

**ME Scholar Kratika Patidar, HOD. Kamlesh Patidar**
Department of Computer Science and Engineering
Jawaharlal Institute of Technology,
Borawan, Khargone (M.P.), India
Kratika.patidar.4@gmail.com,trincharge@gmail.com

***Abstract*-Individuals utilize online media for amusement, bringing data, news, business, correspondence and some more. Not many such online media applications are Facebook, Twitter,WhatsApp, Snapchat, etc. Twitter is one among the miniature publishing content to a blog site. We are utilizing Twitter predominantly because it has acquired a great deal of media consideration. The content composed is alluded to tweets, where an average person can tweet or compose their hearts out. We would be bringing immediate reactions from the general population also. Thus, the information is all the more ongoing. The initial step is to get the tweets on a specific plan utilizing python language code, followed by the cleaning cycle; at that point comes the making of the sack of words. Later these packs of terms are given as a contribution to the calculations. Finally, after preparing the estimates, we will be getting the sentiment of general society on that system.**

***Keywords*- Twitter, Decision Tree, Sentiment Analysis, Supervised Learning, Random Forest, Support Vector Machines.**

## I.INTRODUCTION

Social media platforms are a craze for everyone in today's scenario.Facebook, Twitter, Instagram, Telegram etc., have been used in a huge amount, and data created due to these social media platforms is vast [1].Sentiment analysis is also termed, in other words as Opinion mining.Opinion mining is also becoming an essential task for companies to manage social media platforms [2].

One wrong message spread over the internet can harm humanity in any of the ways.So, we need to apply opinion mining here to get the user's actual mood it may be happy, angry, etc. Also, the behaviour and attitude of the user needed to be understood. Sentiment analysis plays an essential role in this Work. Reactions of users can be Analyzed for various fields such as politics, finance, sociology and economy [3].

The vast amount of data is found on the internet, which is in an unstructured format. In this present data, almost 80% of the data is unstructured, which is needed to be formatted in required format [4].

The structuring of the unstructured data is required as the formless.Data becomes difficult for analysis, and accuracy with such data becomes very less. Sentiment analysis is a type of analysis that works of text placed on the network [5].Twitter is a blogging site which gives the user to share their thoughts, and this site is used by a huge number of users [6].The thoughts or tweets done on any particulartopic can be useful for makingan opinion. This opinion can be of any type that can be used for any purpose.

Some of the major areas are politics, brands, campaign etc [7]. The tweets made on such topics can be useful for analysis. Opinions can be useful when it comes to taking any decision regarding any particular topic. So, this paper presents the Work done for the sentiment analysis of the data taken from Twitter. The paper does the sentiment analysis after cleaning the dataset. For cleaning, we have used TF-IDF, and then this data has been used for the implementation of the classifier for the classification of the sentiments.The dataset used here is taken from the Kaggle website.

Fig 1. Some of the social media sites.

## II. LITERATURE REVIEW

Here authors have done sentiment analysis for the tweets made on donations, fundraising or some charities. They have implemented natural language processing toolkit to

© 2021 IJSRET

check tweet sentiment, which is neutral, positive or negative [8].

The sentiment analysis for any area helps in getting the information of that area. Here the authors have used machine learning algorithms for the sentiment analysis. They implemented an unsupervised learning approach for this Work. The algorithm here used by them is the rule-based or lexical approach. They have taken tweets having hashtags as #haryanaassemblypolls, #theskyispink, etc. [9].

Social media have become an open platform to share personal thoughts, and the shortest form to share is provided by twitter. 90% of peoples share their thought on social media platforms. Here authors have implemented sentiment analysis on statistical tools by use of R programming. Some of the feelings analyzed in this Work are disgust, trust, surprise, anticipation etc. [10].

In this paper, authors have done sentiment analysis for the Twitter data. For this, the authors have used TFIDF along with natural language processing. They have implemented the proposed approach, and the accuracy achieved with the NLP technique is 85.25% [11].

Sentiment analysis has been performed in various ways earlier. But here the authors haveimplemented a sentiment analysis of twitter dataset on real-time. The authors have used 2 algorithms for this purpose that us simple voter and naïve bayes. Among both algorithms, Naïve Bayes algorithm has proved to be more efficient [12].

## III. PROPOSED ALGORITHM

Here we have presented the approach given by us which is used for the task of sentiment analysis. The overall Work is carried out in four stages that are data collection, preprocessing of the data, computation of the features and the classification of the sentiments has been done in the very last stage. A flowchart for the Work that is carried out in this paper has been given below.

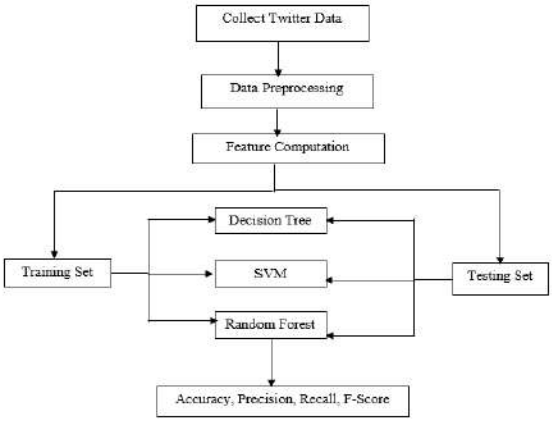


Fig 2. Flowchart for the proposed system.

**1. Data Preprocessing:**
Text mining needs preprocessing of the data on a necessary basis in the overall procedure of the analysis. The data provided to the algorithm as an input makes the basis for the accuracy obtained after a comprehensive evaluation. This preprocessing of data helps remove the stopwords, hashtags, and various languages can be removed through this step.Making whole text in lower case, removing Unicode and obtaining the required data format has been done in this step. The dataset should be free of tags, URLs, along with this space and tabs also need to remove.

**2. Computation of Features:**

- Next step after the data preprocessing is to obtain features from the twitter data, and the representation of such data in vector form is to be done.
- This textual information is converted into sectors to represent this data in numerical format. This Work is useful in machine learning algorithms. If we see the ML can't understand text very well and numerical representations are more feasible.
- We have used TF-IDF for the vectorization of the dataset.

**3. TF-IDF:**
Term frequency and inverse document frequency are the terms that are used widely for retrieving the information from the textual dataset by making use if weighs. Weight calculation of any particular term in text is done by the multiplication of the TF with IDF.

In this procedure any keyword is founded by getting the importance of that particular word and the number of occurrences of that word is defined by the TF-IDF algorithm. One more thing it does is corpus which means that the similar word to the keywords found. Let we have any document d and weight is WT of any term t in that document.
WT, d = TF (t), d log (N/DF(t))
Here, TF(t) is the number of times any term occurs in a document, in whole corpus number of documents present in N, number of total documents that have that term present in it DF(t).

**4. Random Forest:**
It is a supervised algorithm for the classification task. Random forest gives higher accuracy if several trees formed in random forest algorithm is large.

The pseudo-code for the algorithm has been provided below.

- Select k out of m features available.
- From these k features calculate d node using the split points.
- Split further node into daughter nodes
- Repeat step 1 to step 3 until we reach i several nodes.
- Repeat step 1 to step 4 for n times to create n number of trees.

© 2021 IJSRET

**5. Support Vector Machine Algorithm:**
SVM is an algorithm which is majorly used for the task of classification. Various data points have been extracted through SVM and mapping were done in high dimensional space with a non-linear kernel function. Algorithm for SVM has been stated with the help of the pseudo-code given below.

**5.1 Inputs:**Provide different datasets for training and testing of the classifier.

**5.2 Outputs:**Determine the accuracy calculated;

- Choose gamma and cost optimally for SVM
- While (condition = true) do
- SVM training for every data point
- SVM testing for every data points
- End while
- Return accuracy

**8. Decision Tree Algorithm:**
Another classifier we have is a decision tree classification algorithm. This algorithm can be majorly used for the textual classification.

**8.1 Input:** training and testing dataset (tweets).
**8.2 Output:** accuracy, precision, F1 measure, recall
**Start**
Pre-processing and data normalization;
For training dataset to do
Features calculation;
Decision tree algorithm;
Classifier building;
**End**
Value is used for particular tweets;
For testing dataset
Analyze accuracy
**End**
(training, testing)
**End**

## IV. IMPLEMENTATION

**1. Requirements:**
The required software and hardware has been stated below.
For the design Python Programming Language, 15.6 in HD WLED touchscreen (1366 x 768), 10-finger multi-touch support. 10th Generation Intel Core i7-1065G7 1.3 GHz up to 3.9 GHz. 8GB DDR4 SDRAM 2666MHz, 512GB SSD, No Optical Drive. Intel Iris Plus Graphics, HD Audio with stereo speakers. HP TrueVision HD camera. Realtek RTL8821CE 802.11b/g/n/ac, Bluetooth 4.2, 1 HDMI 1.4, 1 USB 3.1 Gen 1 Type-C, 2 USB 3.1 Gen 1 Type-A. The Python Programmingwas run on Windows 10 64-bit Operating System platform. The python library was used during implementation like NumPy, Pandas, Matplotlib, SciPy, Scikit-Learn, PyTorch, Seaborn, XGBoost, Plotly, TensorFlow,Keras,Seaborn, TextBlob,Stanford CoreNLP, Gensim, and Afterword.

**2. Data Collection:**
Here, we have explained the evaluation carried out by our proposed methodology to analyze twitter data. Our experimental analysis is done on the kaggle dataset. Dataset has been obtained from the challenge given by the school of AI (Artificial Intelligence). We have used a portion of 4000 tweets for training purpose and 10000 tweets for testing purpose [13].

**3. Data Set Processing:**
This section provides the details of the experiments that we have performed to analyze 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 tweets for testing purpose. Data set link mention below. https://www.kaggle.com/mcdonalds/nutrition-facts, details figure3 and figure4.

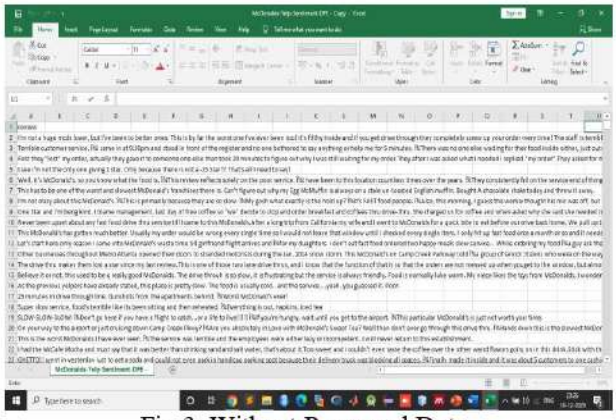

Fig 3. Without Processed Data.

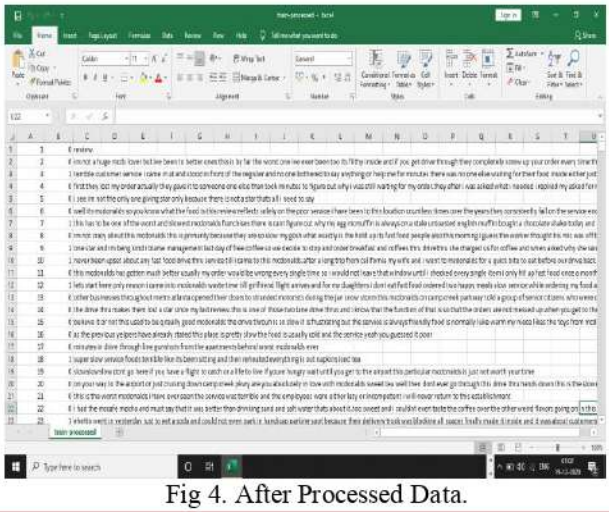

Fig 4. After Processed Data.

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

## V. RESULT ANALYSIS

Performance of each procedure is done no. of times and results have been calculated using different classification procedures on a Twitter dataset. Figures have been drawn using otherproducts as calculated.

Our proposed method evaluated in the below parameters:

- Recall
- Precision
- Accuracy
- F1-Score

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) |

$$recall = \frac{tp}{tp+fn} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$precision = \frac{tp}{tp+fp} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

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

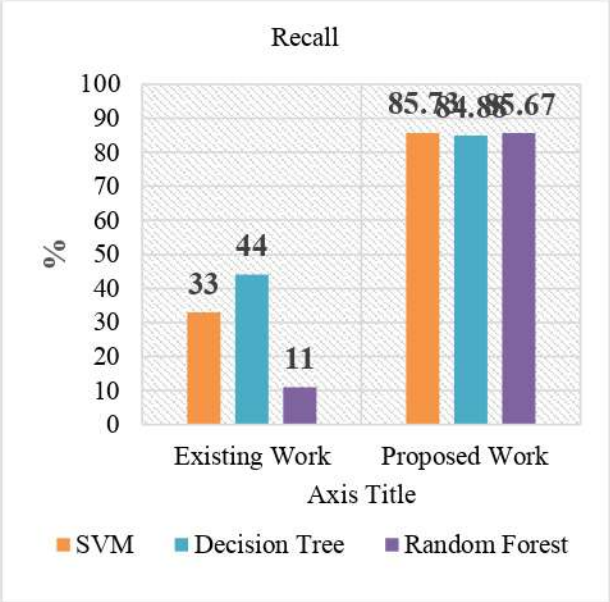


Fig 5. Recall Figure between Existing Work and Proposed Work.

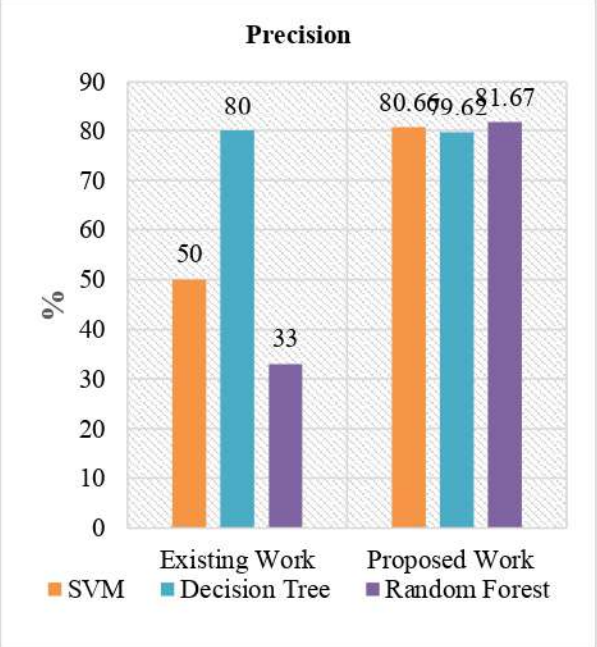


Fig 6. Precision Figures between Existing Work and Proposed Work.

In Figure5 we calculate recall value for all algorithms Existing Work [14] 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 number of positive class expectations that have a place with the positive class.Theprecision is calculated using equation 2.

© 2021 IJSRET

In Figure6 we calculate a precision value for all algorithms Existing Work [16] and Proposed Work and results shown with a diagram's help. 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 (True Positives + True Negatives) to the absolute Test Dataset. The accuracy is calculated using equation 3.

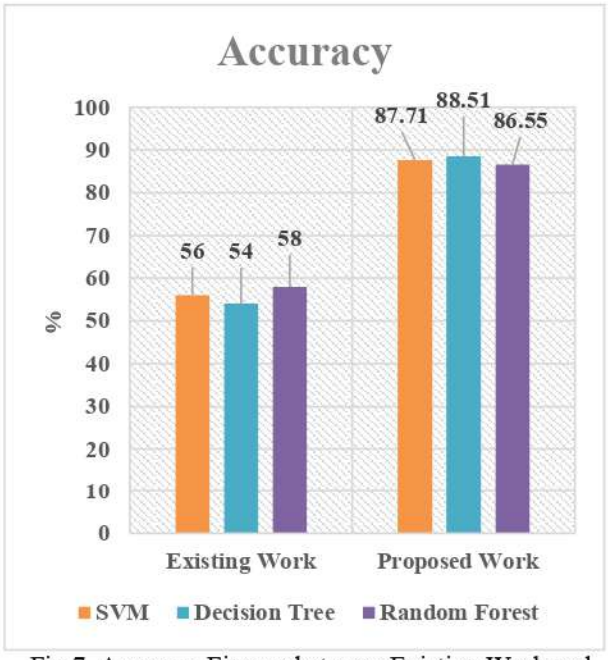


Fig 7. Accuracy Figures between Existing Work and Proposed Work.

Figure7 calculates a precision value for all algorithms Existing Work[14] and Proposed Work and results shown with a diagram's help. We find that the proposed approach shows truer positive rate compare to collaborative and content-based strategies.

**4. F1_Measure:**

The F1 measure states the accuracy of the test; 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 in Figure4.

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

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.

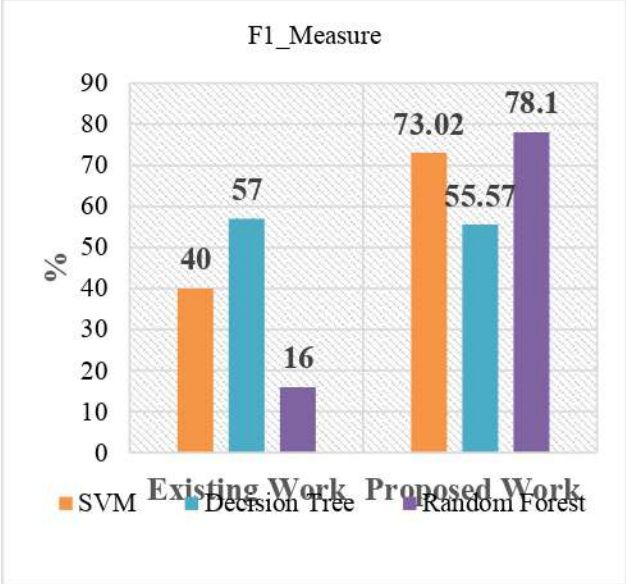


Fig 8. F1_Score Figures between Existing Work and Proposed Work.

## VI. CONCLUSION

Sentiment analysis is necessary to do on social media platforms to stop any spread of hatred messages or issues. We have proposed our Work to do the sentiment analysis of tweets made by users over any particular topic. We have done our data cleaning with the help of available TF-IDF algorithm. The dataset used here was of McDonald's and KFC customers. More than 14000 tweets were analyzed, and from these 4000 tweets were used for training and remaining 10000 for testing. The results were carried out for three different SVM algorithms, Decision Tree and Random Forest, among which best accuracy was obtained for the decision tree algorithm, which valued 88.51%.

## REFERENCE

[1] L. Mandloi and R. Patel, "Twitter sentiments analysis using machine learning methods,"2020 Int. Conf. Emerg. Technol. INCET 2020, pp. 1–5, 2020, doi: 10.1109/INCET49848.2020.9154183.

[2] V. Prakruthi, D. Sindhu, and S. Anupama Kumar, "Real-Time Sentiment Analysis of Twitter Posts,"Proc. 2018 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2018, pp. 29–34, 2018, doi: 10.1109/CSITSS.2018.8768774.

[3] C. W. Park and D. R. Seo, "Sentiment analysis of Twitter corpus related to artificial intelligence assistants,"2018 5th Int. Conf. Ind. Eng. Appl. ICIEA 2018,pp.495498,2018,doi:10.1109/IEA.2018.8387151

[4] S. Dhawan, K. Singh, and P. Chauhan, "Sentiment Analysis of Twitter Data in Online Social Network,"Proc. IEEE Int. Conf. Signal Process.

Control,vol.2019,October,pp.255259,2019,doi:10.1109/ISPCC48220.2019.8988450.

[5] Kusrini and M. Mashuri,"Sentiment analysis in twitter using lexicon-based and polarity multiplication, "Proceeding- 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019, pp. 365–368, 2019, doi: 10.1109/ICAIIT.2019.8834477.

[6] R.Wagh and P.Punde, "Survey on Sentiment Analysis using Twitter Dataset,"Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018, no. Iceca, pp.208–211, 2018,doi:10.1109/ICECA.2018.8474783.

[7] S. A. El Rahman, F. A. Alotaibi, and W. A. Alshehri, "Sentiment Analysis of Twitter Data,"2019 Int. Conf. Comput. Inf. Sci. ICCIS 2019, 2019, doi: 10.1109/ICCISci.2019.8716464.

[8] A. Shelar and C. Y. Huang, "Sentiment analysis of twitter data, "Proc.- 2018 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2018, pp. 1301–1302, 2018, doi: 10.1109/CSCI46756.2018.00252.

[9] S.Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule-Based Approach: A Case Study,"ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir., pp. 537542,2020,doi:10.1109/ICRITO48877.2020.9197910.

[10] S. Saini, R. Punhani, R. Bathla, and V. K. Shukla, "Sentiment Analysis on Twitter Data using R,"2019 Int. Conf. Autom. Comput. Technol.Manag. ICACTM 2019,pp.6872,2019,doi:10.1109/ICACTM.2019.8776685.

[11] M. R. Hasan, M. Maliha, and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data,"5th Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng.IC4ME22019,pp.14,2019,doi:10.1109/IC4ME247184.2019.9036670.

[12] A. S. Al Shammari, "Real-time Twitter Sentiment Analysis using a 3-way classifier,"21st Saudi Comput. Soc. Natl. Comput. Conf. NCC 2018, pp. 1–3, 2018, doi: 10.1109/NCG.2018.8593205.

[13] https://www.iflexion.com/blog/sentiment-analysis-python.

[14] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data,"2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

© 2021 IJSRET