

# Twitter Data Analysis using Machine Learning Classification Techniques

M. Tech. Scholar Nisha Kumawat, Assistant Professor P. Kumar Choure

Department of Computer Science & Engineering  
Swami Vivekanand College of Engineering, Indore  
nishakumawatit@gmail.com, Pkumarchoure@svceindore.ac.in

**Abstract**-Sentiment analysis is the computational study of opinions, sentiments, evaluations, attitudes, views and emotions expressed in text. It refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive or negative sentiment. Sentiment analysis over Twitter offers people a fast and effective way to measure the public's feelings towards their party and politicians. The primary issue in previous sentiment analysis techniques is the determination of the most appropriate classifier for a given classification problem. If one classifier is chosen from the available classifiers, then there is no surety in the best performance on unseen data. So, to reduce the risk of selecting an inappropriate classifier, we are combining the outputs of a set of classifiers. Thus, in this paper, we use an approach that automatically classifies the sentiment of tweets by combining machine learning classifiers with lexicon-based classifier. We physically mark a broad Twitter dataset of 100,000 tweets and perform low quality content discovery continuously dependent on the described noteworthy highlights and word level investigation. The consequences of our exploration demonstrate that our technique has a high precision of 0.9170 and a decent F1 of 0.9460 dependent on an irregular woodland classifier with continuous execution in the discovery of low-quality substance in tweets. Our work in this manner accomplishes a positive effect in improving client involvement in perusing internet-based life content.

**Keywords**-Sentiment Analysis; WordNet; SentiWordNet; Word Sense Disambiguation; Machine Learning Methods; ensemble Approach.

## I. INTRODUCTION

Some social networking applications like Facebook, Twitter, etc are becoming so popular nowadays. Opinion mining is also another term used for this task. Opinion mining comes under the vision of data mining and computation linguistics. The main purpose of applying opinion mining is to detect the behavior, mood, and attitude of the user. As social sites took place in our life, sentiment analysis is used to study the reactions of users in various fields like as politics, economy, finance and sociology.

The data found over social networks is unstructured data. This data is approx 80% of the unstructured data that is present in the world. Because the data is unstructured it is typical for analysis and gains useful information from the available data. Sentiment analysis is the method to read the opinions of the user by textual mining placed over the network[1]. Twitter is a microblogging site that is having a big number of small size utilizes that is for marketing, social networking[1]. When it comes to getting opinions of peoples over any topic then this data is useful for making decisions. For example, various political parties may use this data to make predictions about their results

based on people's tweets. Opinions of peoples are useful when it comes to making any decision. Nowadays to get perspective of peoples Over different issues and to get the conclusions to form their views have become a meaningful way to gain information over any topic. Here in this paper, we have done the sentiment analysis of twitter data.



Figure 1 Social media for sentiment analysis [1].

## II. RELATED WORK

**Sahar A. et al. [2]** The authors proposed a model to work on real-time data produced over twitter. The proposed model varies from the earlier work done. In this area as it uses both the supervised and unsupervised algorithms of ML together. To show the popularity of the two shops Mcdonald and KFC the data was collected for these two.

Results show that McDonald's is more popular as compared to KFC. In McDonald's, 7000 tweets are taken for testing and the author has got 2184 positive, 1589 negative, and 3227 neutral tweets.. In KFC, 7000 tweets are taken for testing and the author has got 2076 positive, 1311 negative, and 3613 neutral tweets.

**Rekha V et al. [3]** Twitter is used mainly as it has got a lot of attention. Direct responses were used from peoples as this makes the data more real. In this paper, hashtags are used to find out the interest of peoples and to sort the topics. The data that is cleaned was applied to following machine learning techniques, i.e. naïve Bayes, random forest, and tailored random forest. Results showed that random forest has more sentiment accuracy as compared to naïve Bayes.

**Sonia Saini et. al. [4]** In this paper, by using R programming language the collected tweets are pre-processed, analyzed and visualized, and for this author presented an open-source approach. R programming tool is used for sentiment analysis. This experiment gives eight different classification feelings and two sentiments. They used approx 3000 tweets to do this classification.

**Nann Hwan Khun et. al. [5]** In this paper, the authors proposed a sentiment analysis based on a visual framework that is used for the analysis of Huawei technologies when the trade war between US-China happened. Sentiment analysis and geographic visualizations were used in the proposed model. The obtained results were useful for different economists and politicians to guess public opinion.

**Sani Kamş et. al. [6]** To do sentiment analysis over twitter data the authors compared different deep learning techniques. Evaluation and comparison of CNN and RNN (LSTM) networks were done. The results show the limitations of CNN and LSTM as their low performance in the field.

**Lei Wang et. al. [7]** Patterns of diffusion for sentiments and the information of text was used to get results and to improve the performance of the sentiment analysis performed over twitter data. The use of both the textual information and diffusion patterns was made as they are related in some context. The algorithm proposed by authors named as SentiDiff to make sentiment prediction over twitter. The given approach achieves PR-AUC increments ranging 5.09% and 8.38%.

**Alaa S. et.al. [8]** Authors presented a system for real-time sentiment analysis on twitter that works online as it provides a real-time system. The graph for polarity of tweets can be carried out by giving input queries. Simple voter and naïve Bayes algorithms are used for the classification of tweets. The proposed 3 way classifier

was analyzed by using the simple voter and naïve Bayes algorithms.

**Ms. Minu Choudhary et. al. [9]** In following, authors used review for mobile companies, they collected more than 5000 reviews from the twitter platform. After that, the applied lexicon approach performs sentiment analysis. From this analysis, they declared that among all brands the Motorola, Samsung, and oppo are more prevalent in the market.

**MerimaČišija et. al. [10]** The authors used the tool Rapid Miner to get the data from twitter, after this they used AYLIEN extension, preparing the dataset and analyzing the sentiments. The sentiments for the president were analyzed which are classified on three bases i.e. positive, negative, and neutral.

**Sheeba Naz et. al. [11]** In this paper, authors present a technique to classify the sentiments on twitter by using the machine learning domain, which uses various textual patterns for example n-grams of twitter data. To analyze the proposed model the performed experiments of 4 sets of n-gram features and 3 weighting techniques. The results examined that unigram features perform very well as compared to others.

### III. PROPOSED WORKING FLOWCHART AND ALGORITHM

In this phase discussion about our proposed work. In the proposed work apply 4 steps. In the first step, we collect twitter data, second step data preprocessing, third feature computation, and lastly use classification technique for tweets analysis.

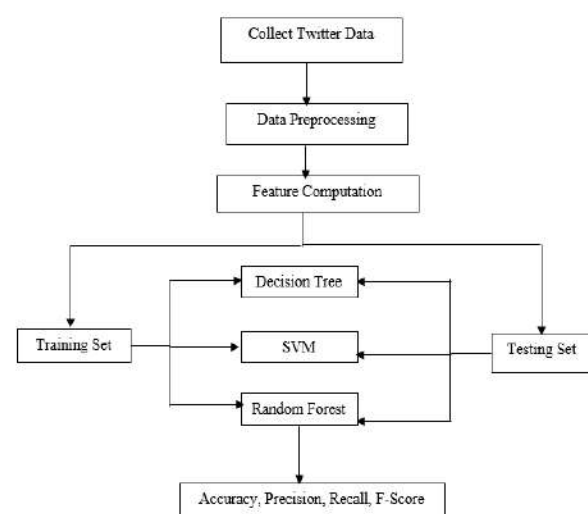


Figure 2 Flow chart of a system overview.

Figure 2 represents about our working process. In the first step we have to collect data from Kaggle then apply TF\*IDF and N-gram process for data preprocessing and

feature computation. The last steps use classification techniques for tweets identification in terms of accuracy, precision, recall, and f-score.

### 1. Data Preprocessing

In data mining the preprocessing of data is considered as an essential procedure. The result's accuracy totally depends on the data set given as an input to the proposed system so the data set needs to be more accurate and having required features. Here in our dataset we are having the linguistic problems as users tweet in their preferred language.

The removal of the stop words, punctuations and the URL is done in the text processing task. Also making the whole text in lower case, removal of Unicode and getting it into a desired format is also been done in this step. The data preprocessing performs some of the basic cleaning of data, removal of noise and the elements that disturbs the original data is removed. And this data is then sent to the next step and the misspelled words are then minimized. For getting efficient results the data should not contain URL, hashtags or the mentioning of peoples, also the tabs or spaces needed to be exchanged with the blank spaces and the quotations with the apex.

### 2. Feature Computation

- After the processing of the data obtained from tweets now the task is to get the required features and the representation of these tweets in the form of vectors.
- The texts are converted into sectors for their numerical representation. This step helps in the training of ML algorithms. In short, the machine cant gets the text and machines well understand the numbers.
- TF IDF does the text vectorization here

**2.1. The term frequency (TF) and the Inverse Document Frequency (IDF)** is one of the widely used methods for the retrieval of information from the texts by using the weighs. Every word in the text is having their own score for the TF and IDF score. The weight of any term is calculated by multiplying the TF and the IDF of that particular term.

The keyword is selected for finding the importance of that word in the document and the occurrence of that term is founded by the TF IDF algorithm. Apart from this it also checks for the more similar words on the internet that is called the corpus. Suppose we have a document presented as  $d$  and the  $WT, d$  id the weight of that any term  $t$  in the whole document.

$$WT, d = TF(t), d \log(N/DF(t))$$

Here:

- Any number occurring in the document is represented by  $TFt$ .

- The all number of documents in which the term is found is presented here by  $DFt$ .
- In the whole corpus, the total document present is given by  $N$ .

### 2.2. N-Gram

Formulation of Problem

Suppose we have any equation  $P(w|h)$ , this is the probability of any word for any history.

Let us consider,

Here,

$w =$  This

$h =$  place is so cool as

an approach named counting relative frequency can also be applied for this probability calculation, in this approach we select a big corpus and we will firm the term place is so cool as, and after this, we will count the terms followed by this.

### 3. Random Forest Method

This is the method that belongs to the supervised learning approach. This algorithm can also implement the classification task. The given large dataset is used for the creation of decision trees in a huge manner. The formula for the prediction is the average improvement of the accuracy predicted. In our proposed approach we are using this algorithm for the tweets classification that is seen on twitter. The tweets are separated on the basis of the patterns seen in the tweets by using the random forest algorithm. The algorithm applied in this classification is given below.

---

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. Nextly we select the  $N$  number of trees that we want to create. Now for the new datapoints 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. SVM

It is another algorithm for the classification task. This creates the decision boundaries, that segregate the n-D area in subclasses. This method used for the selection of the next data points that are helpful in the future. A hyperplane is a term used for the decision boundaries found. We have used the Kaggle dataset and the SVM is best for the text classification and we are doing the same task for tweets. The algorithm is given below:

---

I/P : 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
The four functions of the kernel are applied here through
SVM that is
(linear, polynomials, Sigmoid and Radial Based Function
(RBF));
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

```

---

## 5. Decision Tree

This is another classification algorithm used in such cases for the given data. This algorithm gives a structure like a flowchart, here internal node us “test” for any attribute. Every branch coming shows the outcomings from the previous test. And the leaf node states the label for the class. Text classification can be majorly done through this algorithm. The dataset is the same as we used the Kaggle dataset for analysis. Here the steps for the algorithm is given as:

---

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

```

---

## IV. IMPLEMENTATION AND RESULT

### 1. Hardware and software used for implementation:

The following configuration is used for the implementation of the algorithms.

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 Programming was 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, StanfordCoreNLP, Gensim, and Afterword[12].

### 2. Collect Twitter Data

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 School 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 dataset, we have taken 100000 tweets for our research. 90000 tweets for training and 10000 tweets for testing purposes [13].

### 3. Result

The experiment is tested on the following [arameters that are essential for any of the analyses.

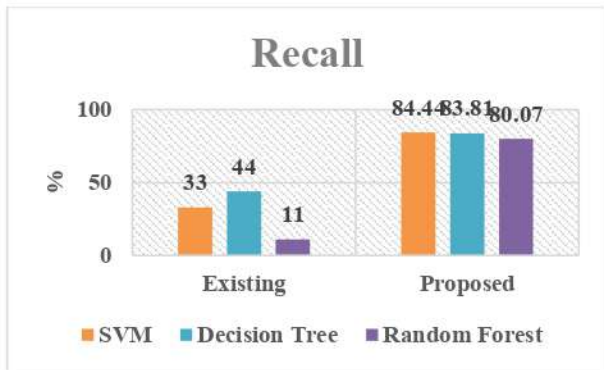
- Recall
- Precision
- Accuracy
- F1-Score

In an actual work compare result in reference paper [14][15].

#### 3.1. Recall

It finds the probability of the finding of the random document in any search retrieval. A recall is that segment, that we have found in the query asked for retrieval.

$$\text{recall} = \frac{|{\text{relevant document}} \cap {\text{retrieved document}}|}{|{\text{retrieved document}}|} \dots\dots(1)$$



Graph 1 Recall graph between Existing Work and Proposed Work.

In graph 4.1 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.

### 3.2. Precision

Precision is the calculation of the probability that the document is found relevant that we have chosen randomly at the start. It is the term that states the portion we are obtaining relevant for the query asked to perform:

$$\text{precision} = \frac{|{\text{relevant documents}} \cap {\text{retrieved documents}}|}{|{\text{retrieved documents}}|} \dots \dots \dots (2)$$

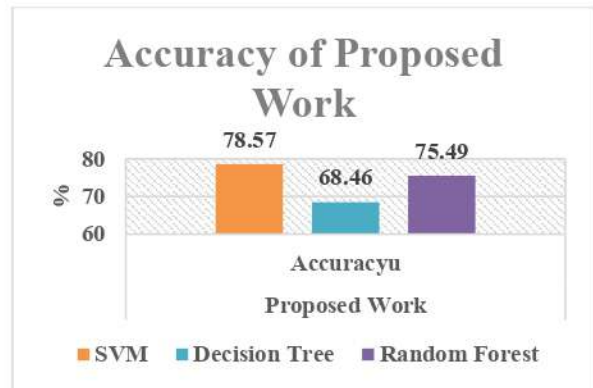


Graph 2 Precision graphs between Existing Work and Proposed Work.

In graph 4.2 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.

### 3.3. Accuracy

It is the percentage of finding the true +ve and -Ve for the overall examination done.



Graph 3 Accuracy graphs between Existing Work and Proposed Work.

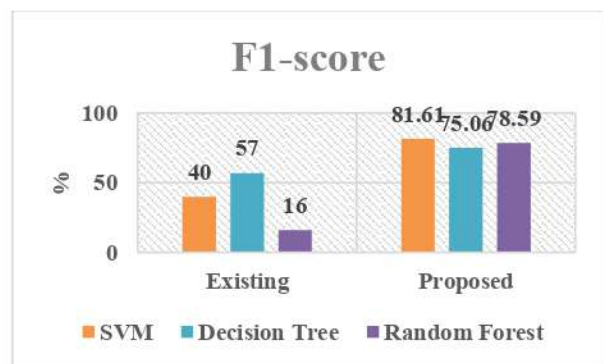
In graph 4.3 we calculate a precision value for all algorithms Existing Work and Proposed Work. And results shown with the help of a diagram. We find that the proposed approach shows truer positive rate compare to collaborative and content-based approaches.

### 4.4. F1\_Measure

The accuracy of the tes 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 graph 4.4

$$F1\_Score = 2 * ((\text{precision} * \text{Recall}) / (\text{precision} * \text{Recall}))$$



Graph 4 F1\_Score graphs between Existing Work and Proposed Work.

Graph 4.4 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.

## V. CONCLUSION

Opinion mining or sentiment analysis comes with a great scope for examining the people's opinion over any topic.

It provides a way to increase the business or to predict the results based on the opinions. Twitter gives a big platform for users to provide their views on any topic. So this data produced over twitter in the form of tweets is used for sentiment analysis. It provides the sentiment of users in a positive, negative, or neutral basis. The various supervised learning algorithms studied in this paper that the results obtained from these algorithms accuracy vary between 68.46 to 78.57%.

## REFERENCES

- [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, FeddahAlhumaidiAlOtaibi 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, RituPunhani, 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 Kaniş and DionysisGoularas, 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, JianweiNiu, 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] Ms. Minu Choudhary and Mr. Prashant Kumar Choudhary, International Conference on Smart Systems and Inventive Technology (ICSSIT 2018), Sentiment Analysis of Text Reviewing Algorithm using Data Mining, 978-1-5386-5873-4/18/\$31.00 ©2018 IEEE
- [10] MerimaČišija, Emir Žunić and DženanaĐonko, Symposium on neural networks and applications 2018, Collection and Sentiment Analysis of Twitter Data on the Political Atmosphere, 978-1-5386-6974-7/18/\$31.00 ©2018 IEEE
- [11] Sheeba Naz, Aditi Sharan and Nidhi Malik, 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), sentiment classification on twitter data using support vector machine, 978-1-5386-7325-6/18/\$31.00 ©2018 IEEE
- [12] <https://www.iflexion.com/blog/sentiment-analysis-python>.
- [13] <https://www.kaggle.com/youben/twitter-sentiment-analysis>
- [14] Sahar A. El Rahman, FeddahAlhumaidiAlOtaibi, Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data". 2019 International Conference on Computer and Information Sciences (ICCIS) 2019, IEEE.
- [15] Rasika Wagh, PayalPunde, "Survey on Sentiment Analysis using Twitter Dataset". Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018). IEEE Conference.