JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Sentiment Classification on Twitter Dataset Using Machine Learning Algorithms

Harshita Tandale
*Computer engineering*
*Modern Education Society's College of Engineering.*
Pune, India.
Harshitatandale14@gmail.com

Pratiksha Shitole
*Computer engineering*
*Modern Education Society's College of Engineering.*
Pune, India.
pratikshashitole2000@gmail.com

Vaishnavi Rubde
*Computer Engineering*
*Modern Education Society's College of Engineering.*
Pune, India
rubdevaishnavi@gmail.com

Mahima Bachhav
*Computer Engineering*
*Modern Education Society's College of Engineering*
Pune,India
mahimabachhav@gmail.com

Prof. Shraddha Khonde
Computer engineering
*Modern Education Society's College of Engineering*
Pune, India.
Shraddha.khonde@mescoepune.org

***Abstract*— NLP (Natural Language Processing) is a subset of Machine Learning that deals with unstructured data from the real world. The 'computational' way of determining whether a given sentence or paragraph is positive, negative, or neutral is known as sentiment analysis. It derives a person's feelings, opinions, and attitudes, among other things. Using sentiment analysis methods to evaluate views and attitudes in Twitter data, which could help businesses better understand how people feel about particular events or products. In this paper, we gathered a dataset of tweets using the hashtags COVID-19 and CORONA VIRUS from Twitter. Sentiment analysis was performed on the Twitter dataset, as well as Supervised Machine Learning Classification methods. The best accuracy of Random Forest is 94.90%.**

***Keywords—Machine Learning, Natural Language Processing,SVM, NB, Sentiment Analysis***

## I. INTRODUCTION

Sentiment analysis is the automated process by which the subjective information underlying a text is identified and extracted. This can be a viewpoint, a judgement, or a sentiment about a particular subject or topic. Polarity detection is the most prevalent type of sentiment analysis, and it is graded as 'positive,' 'negative,' or 'neutral.' Consider the following sentence: "The situation in COVID-19 is not good." This would be automatically classified as Negative by a sentiment analysis model. Sentiment analysis is a sub-field of Natural Language Processing that has gotten a lot of interest in recent years due to its numerous exciting business applications. Twitter is an excellent social media site for people to voice their opinions on various campaigns, events, and other topics. Feeling analysis is particularly beneficial for social media monitoring since it provides qualitative insights in addition to the number of likes or retweets. Around 80% of the world's digital data is unstructured, with social media data accounting for a major chunk of that. Sentiment analysis systems use machine learning and natural language processing to intelligently arrange unstructured text data. Sentiment analysis algorithms can use data samples to learn how to detect the polarity of Tweets in real time. All you have to do is teach sentiment analysis algorithms to recognize emotion in tweets, and they'll take care of the rest More people have reacted to the COVID-19 crisis on social media sites such as Twitter and Facebook, resulting in a massive volume of data being collected.

This data will need to be analyzed in order to gain some insights. Sentiment Analysis is a technique for calculating the polarity of tweets and visualizing the reactions of individuals on COVID-19. In this study, we retrieved COVID-19-related tweets from Twitter and used various sentiment analysis approaches. Machine learning techniques were used to calculate the accuracy of each classifier after the sentiment score was calculated, and a comparison study was conducted.

## LITERATURE SURVEY

If you've spent enough time in the tech business, you've probably heard of sentiment analysis. It's a strategy for judging if a piece of information (typically text) conveys a good, negative, or neutral impression of the subject. Depending on the level of intricacy necessary, sentiment analysis can be performed at three different levels.

### *A. Document Level*

This is the simplest type of sentiment classification, in which the entire document of opinionated text is treated as a single piece of data. It is assumed that the document only has an opinion on a single object (film, book or hotel). If the paper contains conflicting ideas about different objects, this technique is not appropriate. The entire document is reviewed for classification, and whether it is positive or negative is determined. Before processing, inappropriate sentences must be removed.

Document-based sentiment analysis has gotten a lot of attention. . There are two ways to classification: supervised machine learning and unsupervised machine learning.

For supervised machine learning with finite classes for classification, a training and text dataset is available. It uses one of the common classification techniques, such as Naive Bayes, K Nearest Neighbours, Maximum Entropy, and Support Vector Machine, to classify the documents. Combining several supervised approaches to machine learning for effective outcomes was used to produce document-based classification for news comments. For the classification of film reviews, the Naive Bayes and Neural Network classifications are merged. They also demonstrated that by combining these two methodologies, the precision of sentiment analysis can be increased to 80.65%.

Data is analysed using an unsupervised method to machine learning in specific studies. In an unsupervised technique, the sentiment Orientation (SO) of opinion terms in a document is estimated. The normally negative material is marked as positive if the SO of these terms is positive. In the most influential studies, the words "bad" and "excellent" were utilised. The semantic orientation of an opinion determines whether it is similar to the positive word "Excellent" or the negative word "bad." The semantic orientation is calculated using the Point Wise Mutual Information approach. By using a lexicon-based method, they were able to characterise sentiments. The document-level polarity of film reviews is assessed using the unsupervised dictionary-based approach (WordNet). The terms of opinion, as well as their polarity, are included in the seed list in this work.

### *B. Sentence Level*

Because each sentence is considered a separate entity, and each sentence may have different opinions, polarity is decided for each sentence.

A statement can be subjective or objective. The truth is included in the objective phrase. As a result, it will have no bearing on the polarity of the review and should be filtered out. The classification of subjectivity / objectivity is a benefit of the sentence level review. Several distinct approaches are studied and evaluated using supervised machine-learning methodology. Depending on the opinion words in the sentence, the sentence might be characterised as positive, negative, or neutral. It mostly focuses on determining how to correctly categorise the text.

### *C. Feature Level*

Feature level sentiment analysis can provide more fine-grained sentiment analysis on some opinion targets and offers a wider range of E-business applications. Based on comparative subject corpora, this paper presents a method for feature-level SA. They got a corpus from Twitter that has been manually categorized as positive, negative, or neutral at the aspect level. The N-gram around approach produced the best results, with a precision of 81.93 percent, a recall of 81.13 percent, and a measure of 81.24 percent.

The following are some of the most significant benefits of Twitter sentiment analysis:

1. ***Real-Time Analysis:*** Analysis of Twitter sentiment is important to track rapid changes in consumer moods, to identify whether concerns are on the increase and to take action before issues escalate.
2. ***Business:*** In the field of marketing, marketers use it to improve their strategies, to understand the feelings of customers towards goods or brands, how people react to their ads or product launches and why consumers don't

buy any Items. Items.

3. ***Politics:*** *In the political field, this is utilised to keep track of political beliefs, as well as to identify ambiguity and inconsistency between claims and government behaviour. This can also be used to predict election results.*
4. ***Scalability:*** It would take hours of manual processing, and as your data grows it would be impossible to scale in case of huge amount of data. We can convert this manual task to automated and gain valuable insights in a veryshort time.
5. ***Consistent Criteria:*** Two members of the same team will interpret the same tweet differently. You can use one setof parameters to evaluate all of your data by training a machine learning model to perform sentiment analysis on Twitter data, so that results are consistent.
6. ***Public Actions:*** Analysis of opinion is often used to track and interpret social trends, to identify potentially harmful circumstances and to assess the blogosphere's general mood.

## II. PROPOSED METHODOLOGY

Following flowchart shows the proposed methodology for this research work.

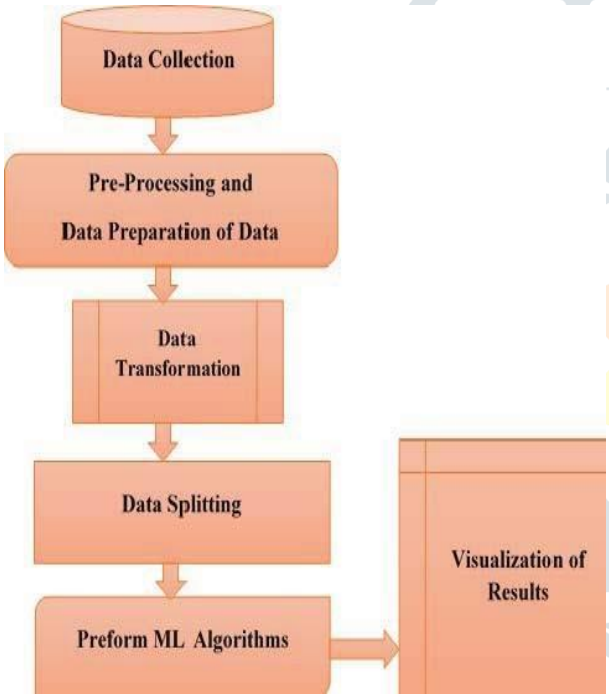


Fig 1. Proposed methodology

### A. Data Collection:

For sentiment classification, kaggle dataset were used in this study.

There are various data sources from which we can obtain various types of datasets, such as text data, video data, audio data, and so on. The dataset used in this paper was downloaded from kaggle and is publicly available. There are 51717 rows (records) and 17 columns in this Twitter dataset (Attributes)

Following Table I. Shows the different attributes in Twitter Dataset:

TABLE I. LIST OF ATTRIBUTES OF TWITTER RATING DATASET

| | | | | |
|---|---|---|---|---|
| url | address | name | online_order | book_table |
| rate | votes | phone | location | rest_type |
| dish_liked | cuisines | approx_cost (for two people) | reviews_list | menu_item |
| listed_in (type) | listed_in(city) | | | |

For the Twitter Dataset, We extracted total 133226 tweets related to two hashtags i.e. COVID-19 and CORONA VIRUS.

### B. Data Pre-processing and Preparation

In this step, we performed different techniques for pre-processing the dataset and made clean for further analysis.

There are various data sources from which we can obtain various types of datasets, such as text data, video data, audio data, and so on. The dataset used in this paper was downloaded from kaggle and is publicly available. There are 51717 rows (records) and 17 columns in this Twitter dataset (Attributes).

### C. Data Transformation

In feature extraction, data transformation is critical. In Machine Learning approaches, features should be numerical. Scaling is a technique for converting all numerical values into a single scale.

- ***Scaling:***

In this step, categorical data has been converted into numerical. Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

- ***Feature Extraction***

When an algorithm's input data is too vast to analyse and is suspected of being redundant, it can be reduced to a smaller collection of features. The selected features should contain the necessary information from the input data, allowing the intended task to be completed using this reduced representation rather of the entire initial data. Positive and negative words were extracted from the Twitter dataset for further study.

### D. Data Splitting

Splitting the dataset into Train and Test sets is critical for training and calculating model accuracy. Training data, which often makes up a specified percentage of an overall dataset together with a testing set, is used to train an algorithm.

### E. Supervised Machine Learning Algorithms

The data in this learning is clearly labelled, and the machine must be trained to predict the target variables for unknown features. Regression and classification are two forms of supervised machine learning techniques.

- ***Logistic Regression***

The data in this learning is clearly labelled, and the machine must be trained to predict the target variables for unknown features. Regression and classification are two forms of supervised machine learning techniques.

- ***Decision Tree***

In the context of classification or regression models, a decision tree creates a tree structure. This cuts down a collection of data into smaller and smaller subsets while also generating a related decision tree incrementally. A decision node can have two or more branches. A ranking or judgement is reflected in the leaf node.

- ***K-Nearest Neighbors (KNN)***

K-Nearest Neighbors (KNN) is one of the most basic Machine Learning algorithms for regression and classification issues. This method takes the data and classifies additional data points based on similarity measurements (e.g., distance function). A majority vote is required for classification to its neighbours.

- ***Random Forest***

The random forest is a decision-making system that comprises of several trees. It employs bagging and feature variability to try to build an uncorrelated forest of trees whose prediction by committee is more reliable than that of any individual tree while creating each individual tree.

- ***Support Vector Machine***

A Support Vector Machine (SVM) is a supervised machine learning model that use classification methods to solve problems involving two groups. SVM models can categorise new text after being given sets of labelled training data for each category. So you're attempting to solve a text classification issue.

- ***Gradient Boosting***

Gradient boosting is a type of machine learning boost. It is based on the notion that when used in conjunction with previous models, the best potential future model reduces total prediction error. The fundamental idea behind this model is to create target outcomes in order to reduce error.

- ***Naïve Bayes***

It's a Bayes Theorem-based classification strategy that assumes predictors are independent. The classifier Naive Bayes believes that a feature's membership in a class has no bearing on any other function.

## III. RESULTS AND DISCUSSION

After pre-processing the Twitter dataset, a sentiment score is calculated, and tweets are classified into three categories: positive, negative, and neutral. Table II follows. This graph depicts the distribution of tweets.

TABLE II. SENTIMENT ANALYSIS ON TWITTER DATASET

| Sr. No | Polarity | Tweet Count |
|---|---|---|
| 1 | Positive Tweets | 88797 |
| 2 | Negative Tweets | 22411 |
| 2 | Neutral Tweets | 21989 |
| | Total | 133206 |

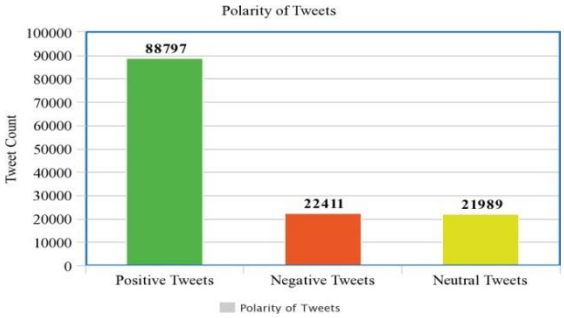


Fig 2. Graphical representation of polarity of tweets

After analysis of Twitter dataset, more people react on Twitter positively about Corona virus. It may be because people might have been helping to increasing confidence of the people and make people positive.

We acquired the following classifier accuracy after deploying different Machine Learning classifiers on both datasets.

TABLE III. CLASSIFIER ACCURACY FOR TWITTERRATING DATASET

After pre-processing the Twitter dataset, a sentiment score is calculated, and tweets are classified into three categories: positive, negative, and neutral. Table II follows. This graph depicts the distribution of tweets.

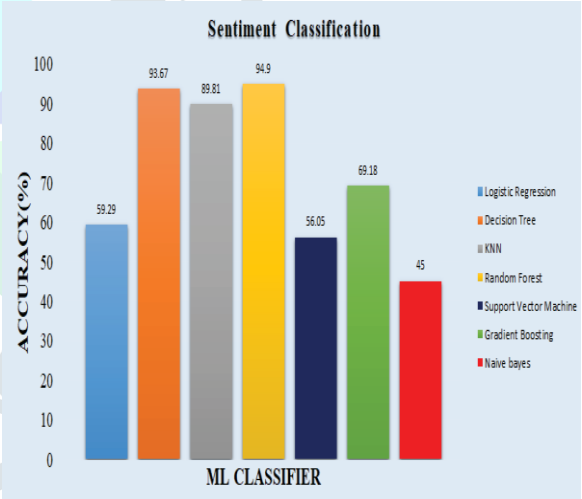


Fig. 3 Graphical representation of accuracy of classfier

## CONCLUSIONS AND FUTURE SCOPE

We can conclude that the Random Forest classifier has the highest accuracy (94.90%) while the Nave Bayes classifier has the lowest accuracy (45%). We can expand the dataset size in the future and use Deep Learning methods. Restaurants Rating System Online is a web-based programme that can be used by the general public to identify appropriate and popular restaurants in any place.

**REFERENCES**

[1] Turney, Peter D., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." arXiv preprint cs/0212032, 2002.

[2] Yan Zhao, Suyu Dong and Leixiao Li, "Sentiment Analysis on News Comments Based on Supervised Learning Method", International Journal of Multimedia and Ubiquitous Engineering, Vol.9, No.7, 2014, pp.333-346.

[3] R. Sharma, S. Nigam and R. Jain, "Opinion Mining Of Movie Reviews At Document Level", International Journal on Information Theory (IJIT), Vol.3, No.3, 2014, pp.13-21.

[4] Raisa Varghese, Jayasree M, "A Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology (IJRET), Vol. 02 Issue 11, 2013, pp. 312-317.

[5] S. ChandraKala and C. Sindhu, "Opinion Mining And Sentiment Classification: A Survey", ICTACT Journal on Soft Computing, Vol- 03, ISSUE: 01, 2012, pp.420-425.
Salas-Zárate, M. D. P., Medina-Moreira, J.,

[6] S. Padmaja and Prof. S Sameen Fatima, "Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC), Vol.4, No.1, 2013, pp. 21- 23.

[7] V. S. Jagtap, K. Pawar, "Analysis of different approaches to Sentence- Level Sentiment Classification", International Journal of Scientific Engineering and Technology, Volume 2 Issue 3, 2013, pp.164-170.

[8] Lagos-Ortiz, K., Luna- Aveiga, H., Rodriguez-Garcia, M. A., & Valencia- Garcia, R.,"Sentiment analysis on tweets about diabetes: an aspect-level approach", Computational and mathematical methods in medicine, Vol. 2017.

[9] Poddar, H., 2020. Twitter Bangalore Restaurants. [Online] Kaggle.com. Available at: https://www.kaggle.com/himanshupoddar/twitter-bangalore- restaurants [Accessed 25 July 2020.

[10] International Journal of Research in Engineering and Technology (IJRET), Vol. 02 Issue 11, 2013, pp. 312-317.
[11] Linguistics, vol. 37, No. 2, 2011, pp. 267-307.
S. ChandraKala and C. Sindhu, "Opinion Mining And Sentiment Classification: A Survey", ICTACT Journal on Soft Computing, Vol- 03, ISSUE: 01, 2012, pp.420-425.