**Int. j. adv. multidisc. res. stud.** 2023; 3(2):595-598

**Received:** 09-02-2023
**Accepted:** 19-03-2023

International Journal of Advanced Multidisciplinary Research and Studies

**ISSN: 2583-049X**

# Sentiment Analysis in Twitter Text by Using Different Types of Machine Learning Techniques

**[1] Tharmini Kiriharan, [2] Ligitha Sakthymayuran, [3] Sivarajah Kiriharan**
[1] Trincomalee Campus, Eastern University, Sri Lanka
[2] Trincomalee Campus, Eastern University, Sri Lanka
[3] Axiata Digital Labs, Colombo, Sri Lanka

Corresponding Author: **Ligitha Sakthymayuran**

**Abstract**
This research seeks to identify the best classifier using Machine Learning (ML) algorithms that can predict the polarity of a comment. The main objective of sentiment analysis is to identify the positive and negative polarities of the social forum text. To conduct this research, we collected sentiment data from the Kaggle dataset and used Natural language Processing (NLP) to classify the emotions from the Twitter text. For that, first preprocess the Twitter text by stemming and cleaning the data by removing Twitter handles, stop words, links, punctuations, numbers, and special characters. Thereafter text tokenization and normalization processes are carried out to the cleaned tweeter text. After that, the frequency word matrix has been created using a count vectorizer. Finally, the accuracy has been calculated by applying different types of classifiers to the word matrix. The accuracy obtained is 95.71% for XGB Classifier, 97.15% for the random forest classifier, 96.28 for logistic regression, 93.24% for the decision tree classifier, and 96.19% for the SVM classifier where this method gets more accuracy than the previous work.

**Keywords:** Twitter, Natural Language Processing, Random Forest Classifier, SVM Classifier, Machine Learning Algorithms

## 1. Introduction
"Since present people not expressing their feeling and thoughts in an open manner compared to the previous time, therefore now sentiment analysis plays a vital role to detect and monitor feelings and emotions in all types of data quickly. Recently, social media platforms including Twitter, Facebook, YouTube, and many others, have surged in popularity. Sentiment analysis is also like opinion mining and it is an approach of natural language processing that identifies the emotion from the text. Its main objective is to extract the human's mood and viewpoint from the documents. As the use of social forum sites has grown, sentiment analysis approaches have begun to leverage the public data on these sites to conduct sentiment analysis research in a variety of sociological fields.

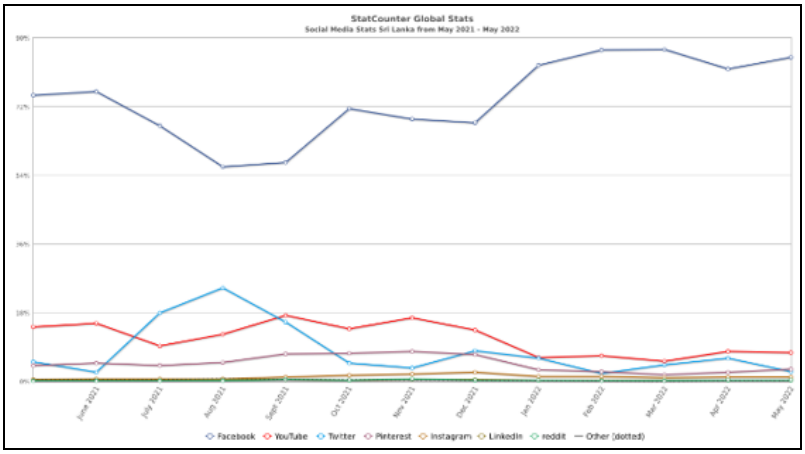


**Fig 1:** Statistical graph of the population of people using social media

Twitter is an in-demand microblogging service where users post status updates. These are called "tweets" and are limited to hundred and forty characters. Tweet to express human opinion on various topics. Compare with other social media text sentiment analyses tweets have main two issues. The first one is the number of misspellings and slang Tweets is much greater than in other domains. Second, Twitter users post their opinions and news in a variety of topics, unlike other social media texts like news, blogs, and other websites tailored to a specific topic.

There are two main techniques are used in sentiment analysis such as symbolic approaches and machine learning approaches. Predefined database machine-learning approaches do not require full emotion, but are simpler than symbolic approaches.

The presence of emoticons and slang spelling mistakes in tweets require a preprocessing step before the feature extraction step. Various classifiers are used for the classification to find the impact on the specific domain with that specific feature vector.

The focus of this research is to produce an accurate result of people's emotional states using Twitter chat data by applying NLP and different types of machine learning algorithms.

Machine learning is a subset of Artificial intelligence (AI) that helps to automatically learn and improve from experience without following explicit programs. Machine learning can be used to solve AI problems and improve NLP by automating processes and providing accurate answers.

The remainder of the essay is structured as follows. The second section briefly describes the previous work done for sentiment analysis in various fields by various scholars. The third section discusses the data and methodology used for sentiment analysis. Then the results and discussion are covered in section four, which is followed by future work and the conclusion in the next steps.

## 2. Literature Review

In the past decades, various text mining approaches have been used. Xing Fang and Justin Zhan [1] solved the problem of emotion polarity categorization, which is one of the key challenges in sentiment analysis. This study makes use of data from Amazon.com's online product reviews. Both sentence level classification and review level classification are investigated in this study. This research has made use of the Scikit-learn program which is an open-source machine learning python-based library. Naive Bayesian, Random Forest, and SVM algorithms were chosen for this classification.

Geetika Gautam and Divakar Yadav [2] are worked on sentiment analysis for the classification of client reviews. This work makes use of Twitter data that has already been classified. In this research, they used three supervised techniques such as nave-Bayes, SVM, and Max-entropy, followed by semantic analysis, which was combined with all three algorithms. They used Python and NLT to train and categorize the models. While SVM combined with the unigram model performs better than SVM alone. The accuracy is increased when the WordNet of semantic analysis is used after the aforesaid procedure.

There has been relatively little research on how emotions are expressed vocally, in contrast to the vast research on nonverbal manifestations of emotion [5]. Because text-based communication tools like instant messaging and email lack nonverbal signals that are generally associated with emotion, understanding the relationship between language verbal expression and emotions is crucially important. It has been proposed that word-based conversation has a reduced ability for emotional interaction since it lacks nonverbal clues.

In one study [14], participants were asked to convey their affinity or disaffinity for a partner in a face to face or computer mediated context. Affinity was also exhibited in both communication scenarios. Verbal signals possessed a higher percentage of relationship information than the face-to-face condition, which is consistent with the social information processing theory's expectations.

Pak *et al* method's [15] method developed a Twitter corpus by autonomously collecting tweets through the Twitter API and labeling those tweets with emotions. That corpus was used to create an N-gram and POS-tag-based sentiment classifier that is based on the multinomial Naive Bayes classifier. There is a chance of error using this strategy because the emotions of tweets in the training dataset are only identified according to the polarity of emoticons. It is also not effective because the training set only consists of emoticon-only tweets.

## 3. Data and methodology

The proposed experiment test with the emotion dataset which is gathered from the Kaggle dataset. This dataset contains 1,600,000 tweets which are extracted using Twitter API. This dataset contains 800,000 positive emotional texts and 800,000 negative emotions texts.

In this research, sentiment classification is used to automatically identify the tweet text based on emotions and label them as 0 for negative texts as shown in Fig 2 and 1 for positive texts as shown in Fig 3.

Thereafter making the processed twitter text more efficient by removing stop words that do not contribute to any future operations. Common words are eliminated in this case, leaving only the special words that reveal the most about the text. After removing stop words twitter text is split into small units called tokens because it is the foundation for developing models further. Normalizing the Twitter text is the next process of transforming Twitter text into a standard form to reduce the noises created by a single word with multiple forms. It is done by using a stemming operation to reduce the word to its root format.

| | target | TweetText |
|---|---|---|
| 0 | 0 | is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah! |
| 1 | 0 | @Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds |
| 2 | 0 | my whole body feels itchy and like its on fire |
| 3 | 0 | @nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there. |
| 4 | 0 | @Kwesidei not the whole crew |

**Fig 2:** Negative text

| | target | TweetText |
|---|---|---|
| 799999 | 4 | I LOVE @Health4UandPets u guys r the best!! |
| 800000 | 4 | im meeting up with one of my besties tonight! Cant wait!! - GIRL TALK!! |
| 800001 | 4 | @DaRealSunisaKim Thanks for the Twitter add, Sunisa! I got to meet you once at a HIN show here in the DC area and you were a sweetheart. |
| 800002 | 4 | Being sick can be really cheap when it hurts too much to eat real food Plus, your friends make you soup |
| 800003 | 4 | @LovesBrooklyn2 he has that effect on everyone |

**Fig 3:** Positive text

| TweetText | Clean_TweetText |
|---|---|
| I LOVE @Health4UandPets u guys r the best! | love health uandpet best |
| im meeting up with one of my besties tonight! Cant wait!! - GIRL TALK!! | meet besti tonight cant wait girl talk |
| witter add, Sunisa! I got to meet you once at a HIN show here in the DC area and you were a sweetheart. | darealsunisakim thank twitter sunisa meet show area sweetheart |
| Being sick can be really cheap when it hurts too much to eat real food Plus, your friends make you soup | sick realli cheap hurt much real food friend make soup |
| @LovesBrooklyn2 he has that effect on everyone | lovesbrooklyn effect everyon |

**Fig 4:** Preprocessed twitter text


**Fig 5:** Positive word cloud


**Fig 6:** Negative word cloud

Then visualize the data using word cloud as shown in Fig 5 & Fig 6 to show which words are the most frequent among the given twitter text.

Finally build the model to apply classifiers to classify the emotions. It was carried out by using Count Vectorizer which is used to convert Twitter text into vector format with the appropriate count of each word that occurs in the Tweets. Thereafter trained the model by dividing the dataset for training and testing with ratio of 70:30.

To train first used the XGB classifier which used DMatrix as an internal data structure to optimize both memory efficiency and training speed.

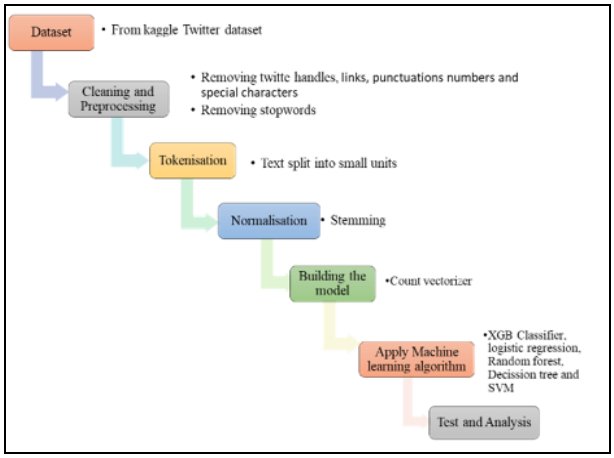


**Fig 7:** Proposed Methodology

After that, we used a random forest algorithm which is one of the supervised learning algorithms that can be used for regression and classification. It develops decision trees for randomly picked data and gets predictions from each tree and gives the best recommendation by means of voting. Fig: 7 illustrates the proposed methodology.

Then the linear classifier logistic regression is applied to classify the data. It is fast and relatively uncomplicated and convenient to interpret the results. Decision Trees are not parametric supervised learning methods for regression and classification. It assists in the development of a model that forecasts the value of a targeted variable by deriving straightforward simple decision rules from data attributes. Finally, classification is performed using a Support Vector Machine (SVM). Here the best accuracy is taken into the consideration by changing the values of parameters and arguments in the classifiers. After applying the different types of classifiers to the processed twitter datasets the overall accuracy is calculated for each classifier.

## 4. Result and Discussions

All of the tests are carried out at Google Collaborator. It enables a fully customized runtime for projects.

For the training and the testing part, the whole dataset is divided into 70 percent and 30 percent, respectively. After training the dataset with several classifiers, the accuracies are tested.

High accuracy is getting for each classifier by changing different parameters and arguments. Finally, the best accuracy is getting for each classifier.

**Table 1:** Overall Accuracy for different classifiers

| Classifier | Accuracy |
|---|---|
| XGB classifier | 95.71% |
| Random forest classifier | 97.15% |
| Logistic regression | 96.28% |
| Decision tree | 93.24% |
| SVM | 96.19% |

As per Table: 1 the overall accuracy obtained for the XGB classifier is 95.71%, for the random forest classifier is 97.15%, for logistic regression is 96.28%, for decision tree is 93.24% and for SVM 96.19%. Among these classifiers, maximum accuracy was obtained from the random forest classifier.

### Future work

In our proposed research, only the handcrafted classification algorithms such as XGB, logistic tree, Random Forest, SVM and Decision tree are used to compare the result. But in the present era, deep learning plays an important role to classify images to get high accuracy. In the future, this research can be extended by comparing with deep learning algorithms.

## 5. Conclusion

To detects sentiments from the tweet text, numerous symbolic and machine learning techniques can be applied. There are certain difficulties when trying to distinguish emotional keywords from tweets with huge number of word. Another challenging part is to deal with slang and misspelled words. The role of the experiment research report is to analyze the sentiment in twitter text. Different classifiers are used to compare the accuracy for the sentiment analysis. Some preprocessing methods on the

dataset are used to get more accuracy and compare vvarious types of classifiers such as XGB classifier, Random Forest classifier, Logistic regression, Decision tree, and SVM classifier.

The overall accuracy obtained for the XGB classifier is 95.71%, for the random forest classifier is 97.15%, for logistic regression is 96.28, for decision tree is 93.24% and for SVM 96.19%. These classifiers are all remarkably accurate. Among these classifiers maximum accuracy was obtained for the random forest classifier.

## 6. References


1. Fang Xing, Justin Zhan. Sentiment analysis using product review data. Journal of Big Data 2.1. 2015; 5.
2. Gautam, Geetika, Divakar Yadav. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.
3. Neethu MS, Rajasree R. Sentiment analysis in twitter using machine learning techniques, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 2013, 1-5. Doi: 10.1109/ICCCNT.2013.6726818.
4. Chung CK, Pennebaker JW. The psychological functions of function words. In K. Fielder (Ed.), Social communication, 2007, 343-359.
5. Fussell SR. The verbal communication of emotion. In S.R. Fussell (Ed.), The Verbal communication of emotion: Interdisciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.
6. De Choudhury M, Counts S, Gamon M. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. In Proc. ICWSM '12, 2012.
7. Kramer A. An Unobtrusive Behavioral Model of "Gross National Happiness". In Proc. CHI, 2010.
8. Niels Rosenquist J, Fowler J, Christakis N. Social Network Determinants of Depression. Molecular Psychiatry. 2011; 16(3):273-281.
9. Paul MJ, Dredze M. You are What You Tweet: Analyzing Twitter for Public Health. In Proc. ICWSM, 2011.
10. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2008, 2011, 2012.
11. De Choudhury M, Counts S, Gamon M. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. In Proc. ICWSM ’12, 2012.
12. Walther JB. Interpersonal effects in computer-mediated interaction: A relational perspective. Communication Research. 2012; 19:52-60.
13. Walther JB, Loh T, Granka L. Let me count the ways: The interchange of verbal and nonverbal cues in computer mediated and face-to-face affinity. Journal of Language and Social Psychology. 2005; 24:36-65.
14. De Choudhury M, Counts S, Horvitz E. Predicting Postpartum Changes in Behavior and Mood via Social Media. In Proc. CHI 2013, to appear, 2013.
15. Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA), 2010.
16. Gautam, Geetika, Yadav, Divakar. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. 2014 7th International Conference on Contemporary Computing, IC3 2014, 2014. 10.1109/IC3.2014.6897213.
17. Singh J, Singh G, Singh R. Optimization of sentiment analysis using machine learning classifiers. Hum. Cent. Comput. Inf. Sci. 2017; 7:3. Doi: https://doi.org/10.1186/s13673-017-0116-3