Comparison of Machine Learning Techniques on Twitter Emotions Classification

S Santhosh Baboo and M Amirthapriya

May 13, 2021

Comparison of Machine Learning Techniques on Twitter Emotions Classification

S Santhosh Baboo[a] and M Amirthapriya[b]

[a]*Associate Professor, P.G. & Research Department of Computer Science, D.G.Vaishnav College, Arumbakkam, Chennai -600106, India*

[b] *Research Scholar, P.G. & Research Department of Computer Science, D.G.Vaishnav College, Arumbakkam, Chennai -600106, India*

Abstract

Social media have become an essential part of our social life nowadays. A huge amount of user reviews and comments shared by the users on social media. Twitter has an excellent growth on social media, and also acts as a platform for business and news. Emotion mining aims to detect, recognize the types of feeling, analyze, and evaluate the human feelings towards any issues or events of their interest. This paper discusses the Twitter text classification using the various machine learning algorithms based on the emotions such as love, anger, anticipation, disgust, fear, joy, optimism, pessimism, sadness, surprise, trust, and neutral. The performance of the classifiers Random Forest, Logistic Regression and Stochastic Gradient Boost is analyzed and the results are being compared. The performance of the classification has been evaluated using TF-IDF, precision, recall, f-measure, and accuracy. From the experimental results, it is observed that Logistic Regression has outperformed by obtaining 77.65% of accuracy than Random Forest and Stochastic Gradient Boosting classifiers.

Keywords

Natural Language Processing, Machine Learning, Twitter, Emotions Classification, Logistic Regression, Random Forest, Stochastic Gradient Descent Boosting.

1. Introduction

Natural language processing is one of the interesting fields of data science. Machine learning is a subfield of Artificial Intelligence, and Machine Learning is the study of computer algorithms which allow computer programs to automatically improve through their experiences [19]. Machine learning models help us in multiple tasks such as Prediction, Classification, Clustering, Object Recognition, Summarization, and Recommender Systems. A machine classifies the spam mails (whether it is a spam or not), reviews (whether it is a positive or negative review), and a search engine which analyzes the user's type, based on their search queries [15].

Text mining is the part of data mining and text mining has several fields like information retrieval, text classification, machine learning, and natural language processing. Emotion mining is the science of detecting, analyzing, and evaluating humans' feelings towards different events, issues, services, and reviews of their interest. Text emotion mining, discusses about analyzing people's emotions based on observations of their text. Text emotion has many applications in routine life such as, customer care services, music and movies recommendations to users, selection of e-learning materials, filtering results of searches by emotion, and diagnosing depression or suicidal tendency [12]. Twitter is one of the most popular and successful platforms where people express their opinions and views about anything through short messages which are known as tweets [16]. Tweets are shown publicly on each user's page and they remain publicly available online for anyone to read and reply to. The public availability of such a huge amount of information about any topic has become a trending field in research nowadays. This paper presents a framework designed to classify the emotions of tweets using the machine learning techniques Random Forest, Logistic Regression, and Stochastic Gradient Boost.

2. Text Classification

Text classification is a trending research part of text mining, where the documents are classified into predefined classes. The classification algorithms are used for predicting a set of items, classes, or categories. Most of the data that are available in social networks like Twitter pages are unstructured. The data extracted from Twitter pages are noisy, unlabeled, occupies more space, and reduce the performance. Since it is difficult to classify the raw data, the noisy data should be removed before the classification process. Preprocessing is the process to remove all unwanted data. Before the text or sentence is fed to a machine, it will need to be simplified first, and this can be done through tokenization and lemmatization. Preprocessing involves four major steps:

- Tokenization
- Stopwords removal
- Stemming
- Lemmatization

Tokenization is the text processing part that helps you to split the large string the pieces of tokens. It means we break down the text into tokens, single, or grouped words, depending on the case. By carrying out a lookup in a pre-defined list of keywords we can ignore stopwords. By doing that, we can free the database space and improve the processing time. There is no universal list of stopwords. The common and frequent terms that are not informative about the corresponding text should be excluded from the text. Stemming is the process of minimizing the derived words to their root form [17]. Lemmatization is a similar activity as stemming, but in lemmatization the base word will have some meaning. Lemmatization transforms some words into their root word.

In this proposed methodology, we are handling feature creation using Count Vectors and Term Frequency Inverse Document Frequency (TF-IDF). Count Vectorization comprises the counting of the number of occurrences each word in a document. Count Vector is a matrix notation of the dataset. In that dataset, each row of the matrix denotes a document from the corpus, each column of the matrix denotes a term from the corpus, and each cell represents the frequency count of a particular term in a particular document. Bag-of-Words is a model that helps you to understand the occurrences of words in a document or a sentence disregarding grammar and order of words. No semantic information is present in the words that we have collected from pre-processing and all the words have the same importance. To overcome these problems, another approach TF-IDF is used. TF-IDF will help you to overcome the problem of a word that has great importance, but presented as in common words in the list.

2.1. Machine Learning Algorithms

Automated text classification is the major research area and this can be achieved through machine learning algorithms. The proposed methodology discussed and analyzed the machine learning approaches Random Forest, Logistic Regression, and Stochastic Gradient Boosting on current trending Tweets. The methods precision, recall, f-score, and accuracy are used here to determine the effectiveness of the classifiers.

2.1.1. Logistic Regression

Logistic regression is a statistical model used for solving the classification problems. Logistic regression estimates the probabilities using a logistic function, which is also referred to as sigmoid function. The hypothesis of logistic regression tends it to limit the function between 0 and 1. This classifier measures the relationship between the categorical dependent variable and one or more independent variables for a given dataset. The dependent variable is the target class, we are going to predict. The independent variables are the attributes that we use to predict the target class [5]. Logistic regression model is also known as Maximum-Entropy classification or log-linear. The logistic regression can be implemented from Scikit-learn library of Python with a class named Logistic Regression.

2.1.2. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. SGD has been applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. SGD is an optimization technique and does not correspond to a specific family of machine learning models. It is only a way to train a model. The advantages of Stochastic Gradient Descent are its efficiency and ease of implementation. The class SGD classifier implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification [21].

2.1.3. Random Forest

Random Forest is a supervised machine learning algorithm and considered as one of the strong methods among all machine learning algorithms. RF is used for the classification and regression problems. The Random Forest is a set of decision trees created by randomly selected training data by the random forest classifier. The final class of the test object is decided by combining the votes from different decision trees. This model works with better accuracy as many decision trees are combined and also it reduces the noise and gives more accurate results. The main disadvantages of a Random Forest algorithm are, complexity, requires more training period, slowness and less effective on real-time predictions since it has a large number of trees [4].

3. Related Work

A survey for sentiment analysis of existing techniques using machine learning algorithms was presented and as a result, the authors had concluded that the Naïve Bayes classifier is insensitive to un-balanced data with more accurate results [11]. Different methods included both lexically-based and supervised machine learning-based classification of identifying emotion in the tweets were performed and evaluated. This evaluation revealed that the ensemble method outperformed all other tested methods when tested on both existing datasets and on the dataset created [13]. A comparison technique for sentiment analysis of political analysis using machine learning algorithms Naïve Bayes, and Support

Vector Machine was carried out using the sentiment lexicon W-WSD, SentiWordNet, Text Blob. The authors resulted that Text Blob results better [1]. Evaluation of classification accuracy for product reviews from Amazon was analyzed using the machine learning algorithms Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine and Logistic Regression based on the training dataset size and the count of n-grams [10]. An automated approach of machine learning algorithms like Naïve Bayes and Support Vector Machine for sentiment analysis of user reviews to rank a product was analyzed [9]. The text classification on Twitter trending topics using the Naïve Bayes achieved the results off f-score as 0.77. The research concluded that Naïve Bayes classifier has the main advantage of taking less time to train the model [6]. Social media posts from Google +. YouTube and Twitter according to their relevance were classified using a reduced set of features and two customized bag-of-words and achieved a final score of 0.68 [2].

4. Proposed Approach

Emotion mining is the process of describing and analyzing the feelings expressed about organizations, products, events, industries, movies, and people on social media [13]. This study focuses on the classification of Twitter text (tweets), based on a unique set of twelve basic emotions such as “love”, “anger”, “anticipation”, “fear”, “disgust”, “joy”, pessimism”, “optimism”, “sadness”, “surprise”, “trust” and “neutral”. For the classification process using the machine learning techniques, we need two sets of data as a training set and a test set. In the proposed methodology, we split the dataset as 75% of data into a training set and 25% of data into a test set.

Figure 1 describes the process flow of the proposed methodology. The model for Twitter text emotion classification involves the machine learning techniques Logistic Regression, Random Forest, and Stochastic Gradient Boosting. The workflow consists of four stages:

1. Data Extraction,
2. Preprocessing,
3. Feature Selection
4. Classification

4.1. Data Extraction

The data are extracted through Twitter API using a Twitter account. The Twitter text dataset is considered for analysis which consist of 3000 required and related data fields collected from March 2020 to May 2020 at a random basis. Based on the emotions of the tweets, the data are labeled into 12 different columns as love, anger, anticipation, fear, disgust, joy, pessimism, optimism, sadness, surprise, trust and neutral in the dataset.

4.2. Preprocessing

Before we train a classifier on the data, we need to preprocess the data (or) text. The preprocessing involves:

- All the text in the tweets is converted to lowercase.
- Words like ‘what’s’ are replaced by ‘what is’ and so on.
- Words with apostrophe ‘I’s’ are replaced by the white space.
- Words like ‘can’t, ‘won’t, ‘we’d, ‘I’ve’ are replaced by ‘cannot’, will not’, ‘we would’ and ‘I have’ respectively.
- Words like ‘n’t’ are replaced by ‘not’.

- Words to such as I'm' and 'We're" are replaced by 'I am' and "are" respectively and so on.
- Words like '\'ll' are replaced by 'will'.
- Tokenization of every single word by a white space.
- All the whitespaces are removed in order to remove the noisy data.

The cleaned data set is now ready and we should remove the missing and null values from the data frame. Finally, preprocessing is done and now the data set is ready for the next level.

4.3. Feature Selection

The preprocessed dataset has many properties and hence the machine learning techniques requires the representation of key features of text for further processing. These key features are considered as feature vectors for the classification task and they provide a numeric representation for the words in the data set. For the classification of emotions, the frequency of words or terms plays an important role.
TFIDF contains the terms frequency that specifies the number of occurrences of the term in a given dataset.

TF-IDF is defined as:

$$\mathrm{TF} = \frac{\text{No. of occurences of a word in a document}}{\text{No. of words in that document}}$$

$$\mathrm{IDF} = \frac{\text{No. of documents}}{\text{No. of documents containing the words}}$$

$$\mathrm{TFIDF} = \mathrm{TF} * \mathrm{IDF}$$

The feature vector transforms words into the numerical value in the integer format.

4.4. Classification

Data classification involves two steps, the first step is the learning step, where a classification model is created from a given dataset. The data from which the classification function or model is learned is the training set. The next step is a classification step, where the model is used to test or predict the class labels for a separate given data. The data set used to test the classifying skill of the learned model is the testing set. Data training and testing are performed by the classification methods using the machine learning algorithms. Our aim is to train the data, in the field of emotion classification whether the emotion is joy, fear, sadness, love, trust, anger, anticipation, disgust, optimism, pessimism, surprise and neutral. The methodology includes the machine learning techniques Logistic Regression, Random Forest and Stochastic Gradient Boosting. A comparison among the performance of the machine learning classifiers has been performed.

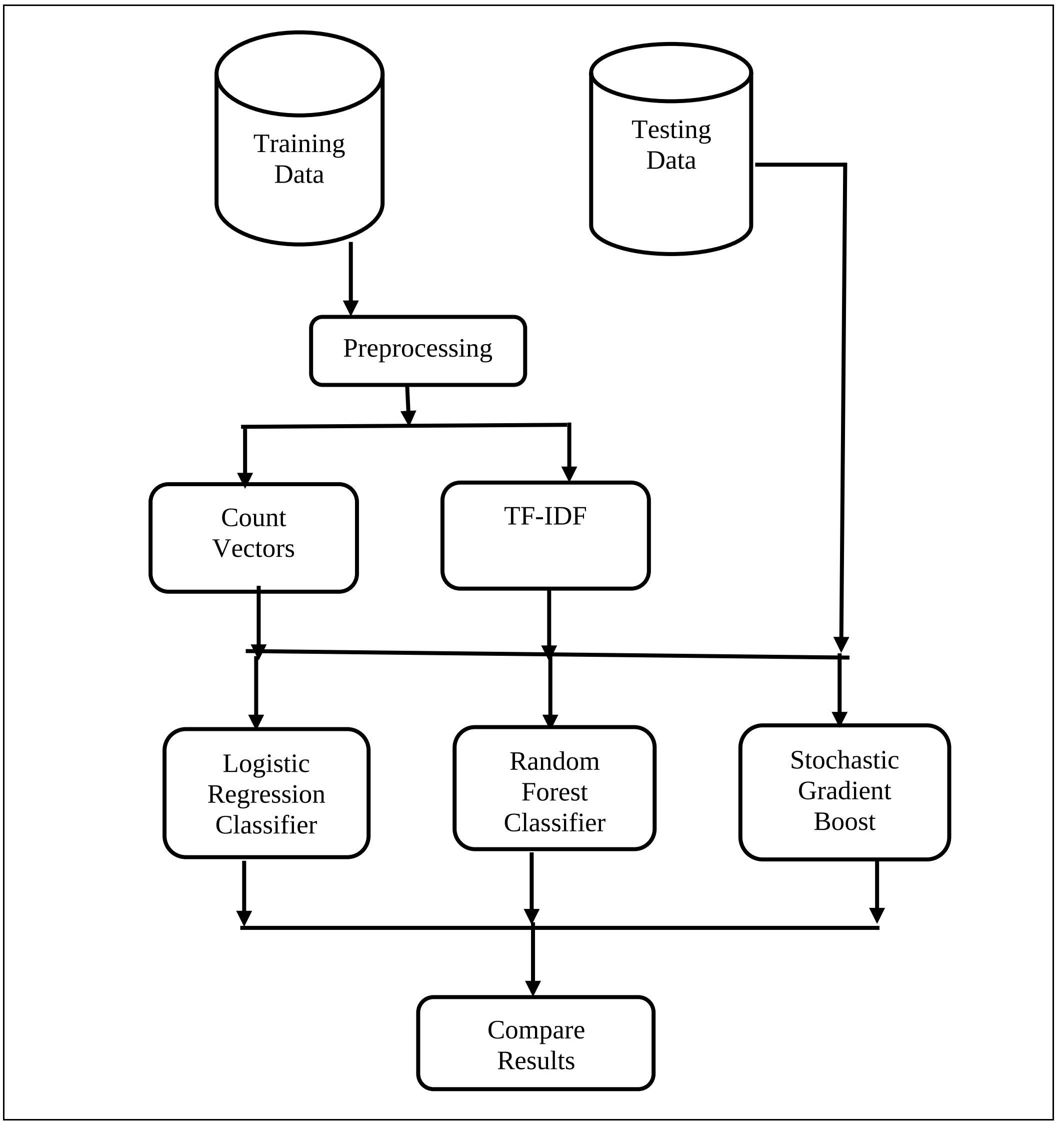

Figure 1: The process flow of the proposed methodology for Twitter emotions classification.

The performance of the classifiers has been calculated using the information retrieval metrics precision, recall, f-measure, and accuracy.

The precision is estimated as (Eq.4):

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \tag{4}$$

where TP is the number of sentences classified to a category correctly and FP is the number of sentences classified to a category incorrectly.

Recall is estimated as (Eq.5):

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \tag{5}$$

where FN is the number of sentences that were not classified at all and TN is the number of sentences marked as being in a particular category and were not. The f-measure is estimated as in (Eq.6):

$$\text{F-measure} = \frac{\text{Precision} \times \text{Recall} \times 2}{(\text{Precision}+\text{Recall})} \qquad (6)$$

The accuracy is estimated as in (Eq.6):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (7)$$

5. Results and Discussion

Figure 2 illustrates the distribution of labeled posts of the training dataset. The vertical axis denotes the posts count and the horizontal axis denotes the text length.

Figure 3 visualizes the distribution of multi-label posts of the training data which denotes the posts that have multiple labels in the training set. The x-axis denotes the index values of the emotions and the y-axis denotes the number of occurrences. The tweets with emotions anticipation, fear, and disgust are high as compared with the other emotions in training data.

Figure 4 visualizes the distribution of multi-label posts of the test data. The x-axis denotes the index values of the emotions and the y-axis denotes the number of occurrences. The tweets with emotions fear are high as compared with the other emotions in test data. The surprise, trust and neutral tweets are remaining low compared with other emotions.

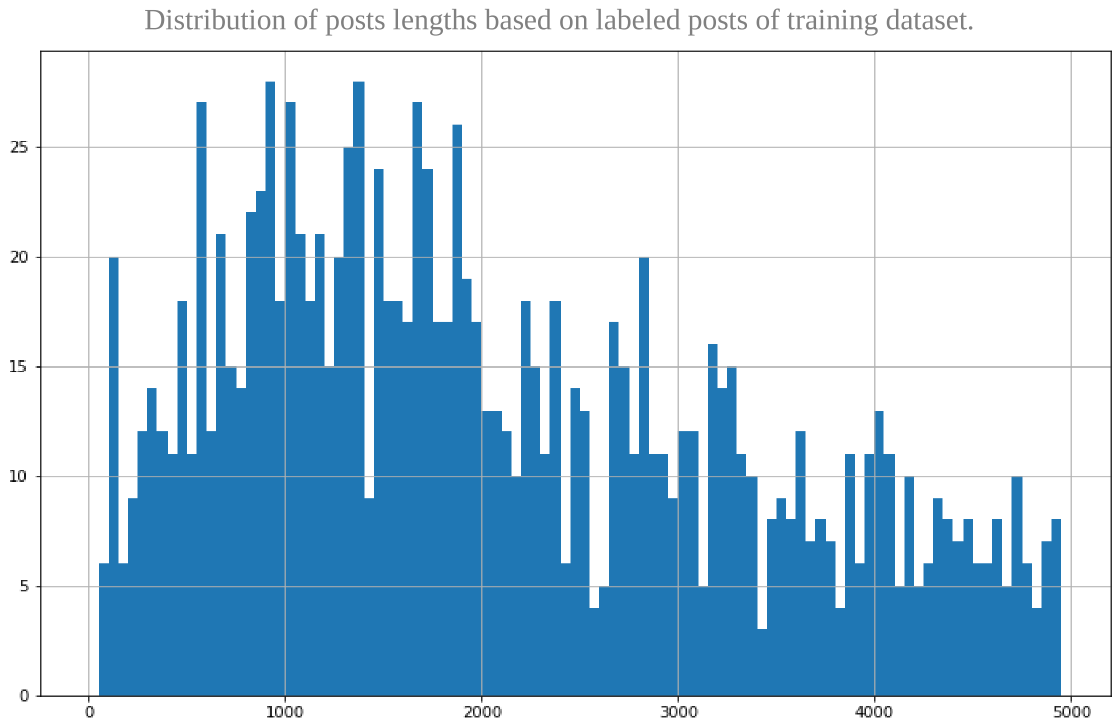

Figure 2: The distribution of text lengths of training dataset.

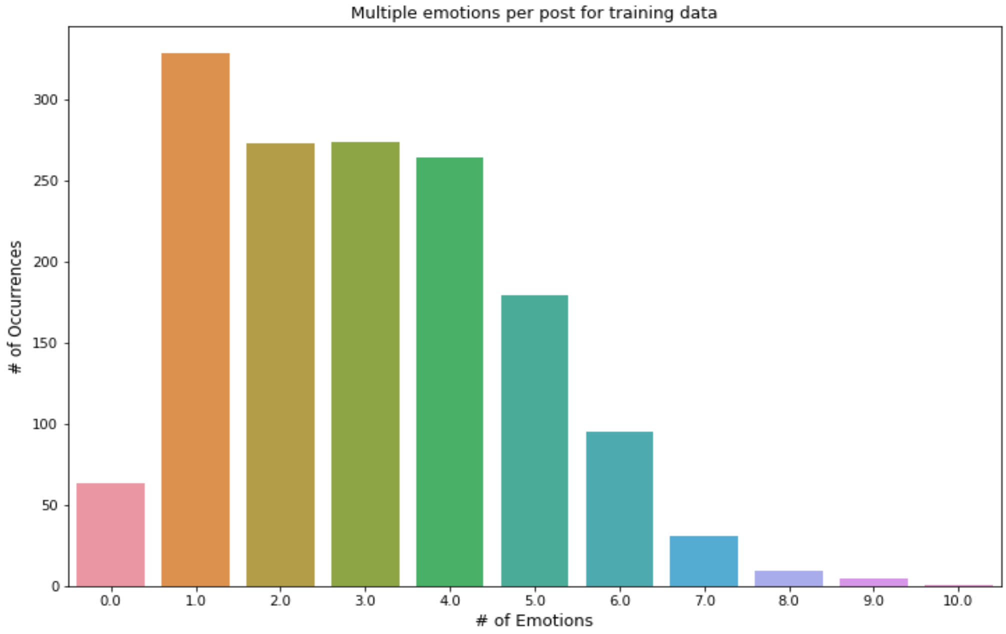

Figure 3: Multiple emotions per post on the training set.

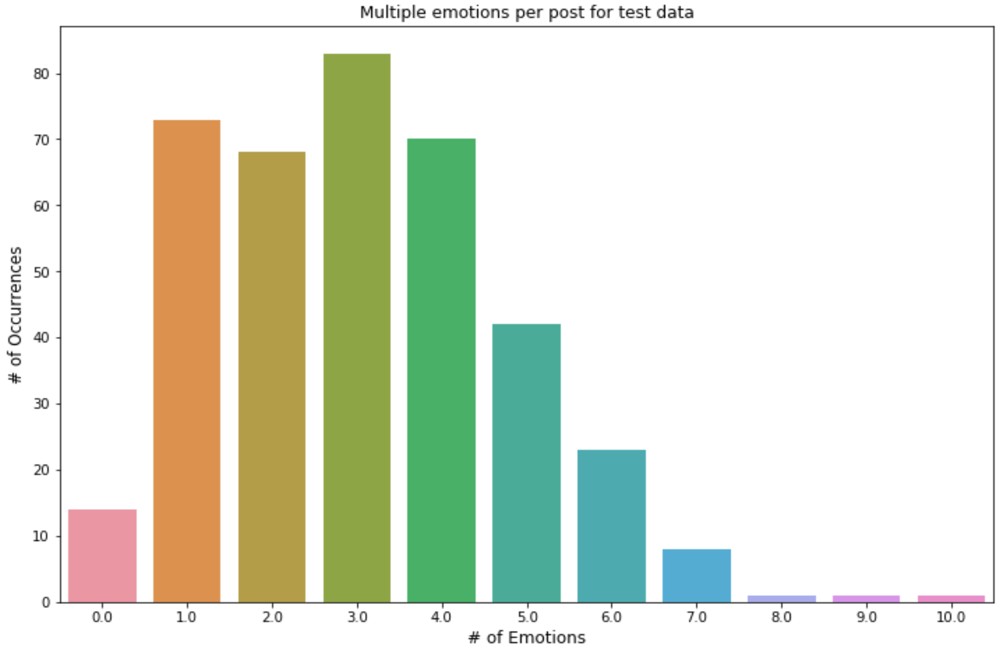

Figure 4: Multiple emotions per post on the test data.

5.1. Logistic Regression Classifier

Logistic Regression Classifier is applied with the proposed TF-IDF features, and evaluation metrics precision, recall, f-measure, and accuracy. Table 1 displays the performance results of the precision, recall, and f-score on various emotions for the Logistic Regression classifier.

Table 1

Performance of Logistic Regression Classifier based on multiple emotions.

Logistic Regression Classifier			
Emotions	Precision	Recall	f-score
Anger	0.7673	0.4132	0.5347
Anticipation	0.6524	0.6513	0.6859
Disgust	0.7326	0.6666	0.6753
Fear	0.6871	0.5538	0.6486
Joy	0.8128	0.0000	0.0000
Love	0.9197	0.0000	0.0000
Optimism	0.7459	0.1546	0.24
Pessimism	0.7192	0.5416	0.5531
Sadness	0.6604	0.0735	0.1360
Surprise	0.8743	0.0408	0.0784
Trust	0.8475	0.05	0.0952
Neutral	0.8983	0.0000	0.0000

5.2. Random Forest Classifier

Random Forest classifier is applied with the proposed TF-IDF features, and evaluation metrics precision, recall, f-measure, and accuracy. Table 2 displays the performance results of precision, recall, and f-score on various emotions for the Random Forest classifier.

Table 2

Performance of Random Forest classifier based on multiple emotions.

Random Forest Classifier			
Emotions	Precision	Recall	f-score
Anger	0.6764	0.0	0.0
Anticipation	0.6176	0.5596	0.5596
Disgust	0.5989	0.0384	0.0740
Fear	0.5213	0.0923	0.1674
Joy	0.8128	0.0000	0.0
Love	0.9197	0.0	0.0
Optimism	0.7406	0.0	0.0
Pessimism	0.6764	0.0	0.0
Sadness	0.6363	0.0	0.0
Surprise	0.8689	0.0	0.0
Trust	0.8395	0.0	0.0
Neutral	0.8983	0.0	0.0

5.3. Stochastic Gradient Boosting Classifier

SGB classifier is applied with the proposed TF-IDF features, and evaluation metrics precision, recall, f-measure, and accuracy. Table 3 displays the performance results of precision, recall, and f-score on various emotions for the Stochastic Gradient Boosting classifier.

Table 3

Performance of Stochastic Gradient Boosting Classifier based on multiple emotions.

Stochastic Gradient Boost Classifier			
Emotions	Precision	Recall	f-score
Anger	0.6764	0.0	0.0
Anticipation	0.5187	0.4082	0.4972
Disgust	0.5855	0.0192	0.0372
Fear	0.4866	0.0205	0.0399
Joy	0.8128	0.0	0.0
Love	0.9197	0.0	0.0
Optimism	0.7406	0.0	0.0
Pessimism	0.6818	0.0083	0.0165
Sadness	0.6363	0.0	0.0
Surprise	0.8689	0.0	0.0
Trust	0.8395	0.0	0.0
Neutral	0.8983	0.0	0.0

5.6. Comparison of the classifiers

The performance of the machine learning classifiers Logistic Regression, Random Forest, and Stochastic Gradient Boost are compared and the result are shown in Table 4. From the results shown in Table 4, the Logistic Regression classifier has outperformed the other classifiers by achieving an accuracy of 77.65%. Figure 6 demonstrates the comparison of the performance of the five classification methods among the classifiers.

Table 4

The comparison of accuracy results of the machine learning classifiers.

Classifier	**Results**
Random Forest	73.395722
Logistic Regression	77.651515
Stochastic Gradient Boosting	72.214795

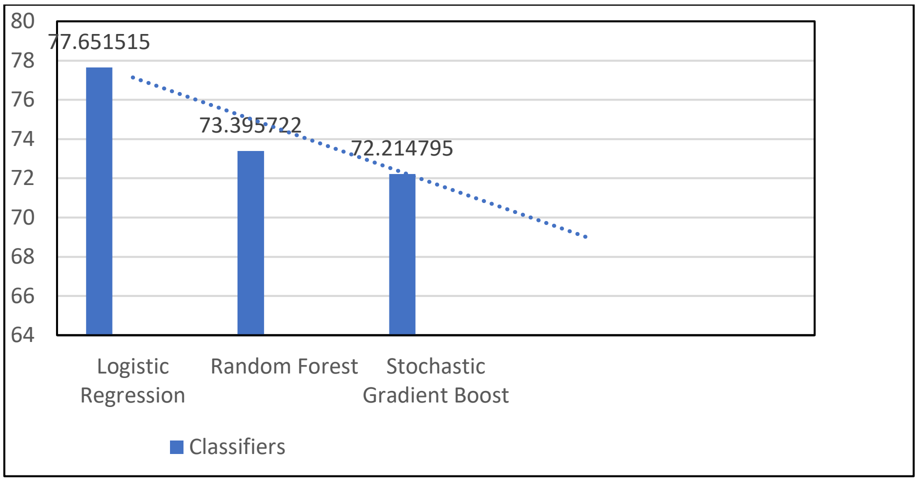

Figure 6: The comparison of accuracy results of the classifiers.

6. Conclusion

This study analyzed the performance of the machine learning algorithms Logistic Regression, Random Forest, and Stochastic Gradient Boosting for Twitter emotions classification. The current trending tweets dataset have been collected through Twitter API. The dataset contains 3000 tweets including all emotions. The results have shown that the Logistic Regression classifier is achieving 77.65% of accuracy. Hence, we conclude that the Twitter dataset with multiple emotions using TF-IDF features, Logistic Regression performs better than Random Forest and Stochastic Gradient Boost classifiers. In future, the analysis of more machine learning algorithms on Twitter emotions classification should be done with the improved dataset, to achieve the best accuracy results.

7. References

[1] Hasan, A., Moin, S., Karim, A., Shamshirband, S. Machine Learning-Based Sentiment Analysis for Twitter Accounts. Math. Comput. Appl. **2018**, *23*, 11. https://doi.org/10.3390/mca23010011.

[2] Alvaro Figueira, M. Sandim, and P. Fortuna. An Approach to Relevancy Detection: Contributions to the Automatic Detection of Relevance in Social Networks, pages 89–99. Springer International Publishing, Cham, 2016. ISBN 978-3-319-31232-3. doi:10.1007/978-3-319-31232-3 9.

[3] Bhagyashri Wagh, Prof. J. V. Shinde, Prof. P. A. Kale, A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques, December 2017, International Journal of Emerging Research in Management & Technology.

[4] Zohre Sadeghian, Ebrahim Akbari, Hossein Nematzadeh, A hybrid feature selection method based on information theory and binary butterfly optimization algorithm Engineering Applications of Artificial Intelligence, Volume 97,2021,104079.

[5] Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc. 1992;41(1):191–201.

[6] D. Irani and S. Webb and C. Pu. Study of Trend-stuffing on Twitter through Text Classification. In Proceedings of 7th Collaboration, Electronic messaging.

[7] Gamallo, P., & Garcia, M. (2014). Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), (SemEval), 171–175.

[8] Gupta M. R., Bengio S., Weston J. (2014). Training Highly Multiclass Classifiers. [ed.] Koby Crammer. Journal of Machine Learning Research. (2014), Vol. 15.

[9] Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1–5).

[10] Pranckevicius, Tomas & Marcinkevičius, Virginijus. (2017). Comparisons of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Baltic Journal of Modern Computing. 5. 10.22364/bjmc.2017.5.2.05.

[11] Sadhasivam, J., & Kalivaradhan, R. B. (2017). Review on Sentiment Analysis a Learners' Opinion. Iioab - Emerging trends in Computer Engineering and Research, 8(2), 298–303.

[12] Shahraki A. G., 2015, Emotion mining from text Master Thesis, Dept. Com. Sci. University of Alberta (Edmonton, AB, Canada). URL: https://era.library.ualberta.ca/items/27ae961f-d9a6-4a5a-9b6f-f180478ea573.

[13] Suboh Alkhushayni, Daniel Zellmer, Ryan Debusk, Du'a Alzaleq. "Text emotion mining on Twitter", IOP SciNotes, 2020.

[14] Medium. URL: https://medium.com/

[15] KDnuggets. URL: https://www.kdnuggets.com/topic/machine-learning

[16] Harjule, Priyanka & Gurjar, Astha & Seth, Harshita & Thakur, Priya. (2020). Text Classification on Twitter Data. 160-164. 10.1109/ICETCE48199.2020.9091774.

[17] Diego Lopez Yse, Your Guide to Natural Language Processing (NLP), 2019.URL: https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.

[18] A, Poornima & Priya, K. (2020). A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques. 493-496. 10.1109/ICACCS48705.2020.9074312.

[19] R. Iriondo, Machine Learning (ML) vs. Artificial Intelligence (AI) — Crucial Differences, 2018. URL: https://pub.towardsat.net/differences-between-ai-and-machine-learning-and-why-it-matters-1255b182fc6.

[20] KDnuggets. URL: https://www.kdnuggets.com/2019/01/solve-90-nlp-problems-step-by-step-guide.html.

[21] Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/modules/sgd.html