International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

COMPARATIVE STUDY ON MACHINE LEARNING APPROACHES FOR TEXT CLASSIFICATION

[1]D Gayathri Devi, [2] Rama Lakshmi
1,2 Assistant professor ,Matrusri Engineering College
[1]gayathridevi.raj20@gmail.com , [2] lakshmi.ramait@gmail.com

Abstract-
Text Classification is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It has thus become a necessity to collect and study opinions on the Web. Of course, there are opinionated publications outside of the Web as well, and many organizations also collect consumer feedback from emails, call centers, and surveys to gauge what their customers think of their products. there are many classifier systems used for predicting the polarity of collected data while using Machine Learning (ML) algorithms. Our project is to essentially evaluate different ML models like Random Forest Classifier, Logistic Regression, and Decision Tree Classifier designed for text classification using different algorithms and analyze the performance of each model. To build our model, we use the Twitter dataset. Further data extraction, processing, and modeling are done on the dataset before using it for training the models. then models are evaluated and compared with other models based on their performance.

Keywords*—text Classification, classifier systems, Random Forest Classifier, Logistic Regression, and Decision Tree Classifier*

Received: 25-07-2025 Accepted: 29-08-2025 Published: 06-09-2025

I INTRODUCTION

In recent times, there has been a significant increase in the number of online customer reviews for products, primarily driven by the rapid growth of online reviews and the surge in social media platforms. These reviews are collected from various sources like Facebook, Twitter, and online forums, and undergo a sentiment analysis process to determine the overall sentiment polarity. This task, which involves extracting opinions or sentiments, is crucial and combines techniques from data mining and natural language processing (NLP) [1].Sentiment analysis, also known as opinion mining, is a computational study focused on understanding people's opinions, sentiments, evaluations, attitudes, moods, and emotions. It is an active area of research in fields such as NLP, data mining, information retrieval, and web mining [2]. The primary purpose of sentiment analysis is to detect whether a given text expresses a positive or negative sentiment. It is commonly employed by businesses to assess sentiment in social data, evaluate brand reputation, and gain insights into customer perspectives [3].There are various models available for performing sentiment analysis, each capable of detecting different levels of sentiment within a sentence. These models take into account different aspects of sentiments and the words used. However, each model has its own advantages and disadvantages, depending on the specific application. It is crucial to determine the efficiency of these sentiment models, particularly in terms of selecting the appropriate model based on factors such as the available data, required accuracy, and application efficiency. This research aims to provide a comprehensive analysis of the performance of various machine learning models for sentiment analysis. Different models, including logistic regression, random forest, decision tree classifiers, and support vector machine (SVM) classifiers, are constructed and evaluated to assess their performance.

I.A Motivation

Text classification is an emerging field of study within text mining analysis. It focuses on

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

organizing and categorizing the vast amount of unstructured text that people generate to express their ideas, opinions, and viewpoints. These opinions can originate from various sources such as the general public, consumers, social media users, athletes, the entertainment industry, and industrial organizations. With the widespread use of social networking services like Facebook, Twitter, and Google Plus, billions of users worldwide contribute to the generation of substantial textual data.

Social media platforms play a significant role in producing diverse forms of text data, including tweet IDs, status updates, reviews, author information, content, and tweet status updates. This abundance of data offers valuable insights into people's perspectives and behaviors. However, due to the unstructured nature of this data, it requires effective text classification techniques to organize, analyze, and derive meaningful information from it. Researchers are actively exploring and developing methods to classify and categorize these vast amounts of textual data for various applications and industries.

II SYSTEM STUDY

II.A.RELATED WORK

Our literature review aimed to explore general online learning algorithms and determine their suitability for our specific use case. Over the past decade, sentiment analysis has been a prominent area of research, rapidly advancing into new domains. A significant portion of this research has focused on developing more accurate sentiment classifiers [4]. The emphasis in sentiment analysis research has primarily been on enhancing the precision of sentiment classifiers, often employing supervised machine learning techniques and various variables.Text classification tasks have been approached at different levels of granularity in natural language processing. Researchers such as Hu and Liu (2004), Kim and Hovy (2004), Wilson et al. (2005), Agarwal et al. (2009), and more recently, at the document level, have explored classification methods for sentiment analysis [Turney, 2002; Pang and Lee, 2004]. However, microblog data, like those found on Twitter, presents unique challenges due to the real-time nature of user comments and reactions to diverse subjects.Several studies have addressed sentiment analysis of Twitter data, including works by Go et al. (2009), Bermingham and Smeaton (2010), and Pak and Paroubek (2010). Go et al. (2009) employed distant learning techniques to gather sentiment data, distinguishing between positive and negative tweets using emoticons. They built models using Naive Bayes, MaxEnt, and Support Vector Machines (SVM), with SVM proving to be the most effective classifier. Their experimentation involved playing with features such as Unigram, Bigram, and parts-of-speech (POS) features, ultimately finding that the unigram model outperformed the others. Bigrams and POS features yielded less satisfactory results.Pak and Paroubek (2010) also utilized distant learning techniques in their data collection process. They trained classifiers using "manufactured" features to analyze Arabic sentiment. Word n-grams were the most commonly used features, employed in training SVM [5, 6, 7], Naive Bayes (NB) [8, 9], and ensemble classifiers [10]. These features are simple but lack semantic depth. To enhance accuracy, word n-grams were combined with stylistic criteria, such as letter and digit n-grams, word and document lengths, and vocabulary richness, after reducing the feature space using the Entropy-Weighted Genetic Algorithm (EWGA) [11]. Additionally, emotion lexicons were incorporated to provide deeper semantic information and improve accuracy. Various lexicons were developed for Modern Standard Arabic (MSA) and dialectal Arabic, including ArabSenti, ArSenL, and SLSA, which were used to train the models [12, 13, 14]. Other factors that could indirectly impact the system were also considered [15].

II.B. EXISTING SYSTEM

There are many existing approaches to text classification. These approaches to sentiment

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

analysis can be grouped into three main categories: Knowledge-based techniques classify text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Some In addition to listing words with obvious affect, knowledge bases often assign arbitrary words a probable"affinity" to specific emotions. Statistical techniques incorporate machine learning components including deep learning, semantic space models, support vector machines, "bag of words" for semantic orientation, and latent text categorization. Deep parsing of the text yields grammatical dependence relations. These Different approaches for analysis of lead to a conflict in the performance of the system where it is being applied. In each approach, the performance is varied based on the dataset length, amount of time for training the model, type of dataset used to train the model, and other factors that may indirectly impact the system.

II.C. PROPOSED SYSTEM

We propose an evaluation system to analyze different Machine Learning Models for text classification, which evaluates the given models for the textual classification based on the performance evaluation factors like training accuracy, Validation accuracy, Confusion Matrix of the models, and f1 scores of each model taken for the performance analysis. The three distinct steps in the performance analysis of the models are as follows: first, the dataset, in this case the Twitter dataset, is prepared for data preparation; next, various models for sentiment analysis are developed; and finally, the models are trained using the dataset that is shared by all the models that will be used in the performance analysis. Then the next step is to evaluate the models based on the performance factors of each model.

II.D.SYSTEM ARCHITECTURE

Fig 1 Module design of system.

III METHODOLOGY

III.A Dataset

The research utilizes the Twitter dataset, consisting of tweets from Twitter users. This dataset, along with other large datasets, provides valuable data for sentiment analysis models. The collected data undergoes preprocessing steps, including null inspection, categorizing sentiments as positive or negative, and analyzing sentiment-related factors. The dataset is transformed into a practical format and divided into groups based on sentiment. Preprocessing includes visualizing and verifying data distribution. The data is assessed and sorted by length taking into account the limits of machine learning models. During preprocessing, unnecessary characters like hashtags are removed, and crucial characters for emotion analysis are modified. The dataset's characters are also standardized in terms of case. As an open-source dataset, user data privacy is not a concern.

III.B. Vectorization

In this methodology, machine learning models are used to classify tweets, but since tweets contain characters that cannot be directly categorized, they are converted into vectors. This vectorization process is necessary to enable the models to analyze and classify the tweets accurately. Various vectorization tools are available, such as gloVe, word2vec, and count vectorization. For this research, the count vectorization method is employed to convert tweets into vectors. The count vectorization technique is applied to the supplied dataset, which involves gathering word frequency and summing up the words in the dataset. This vectorization step allows the machine learning

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

models to process and classify the tweet data effectively.

III.C. Collecting the hashtags

Hashtag extraction is one of the pre-processing procedures before the data set is actually used by the models for training.The data is taken from the tweets in the data set during this preprocessing operation, and the hashtags are then divided into two categories: racist and sexiest hashtags, and non-racist and non-sexist hashtags.

III.D. Tokenising the words

Tokenization is a technique used to break up text into "Tokens," which are words, symbols, and other significant elements.Whitespace characters can be used to divide up tokens.The normalisation procedure involves finding the abbreviations that are present in the tweets, following which the abbreviations are replaced with their full meaning, for example, "OMG" is changed to "Oh My God" [17].If a word appears more than once, its specific meaning will be eliminated.Also, remove the Stop words, Http links, and slang terms like @, RT, etc.breaking a word into tokens that can be combined to make a Unigram.Remove stop words from the list of Unigram words while constructing new words.This phrase is used to assess Chi-squared.The classifier accepted these Chi-squares. They are receiving the list of unigrams. Similar to Bigram and Trigram, the Chi-squared result is evaluated. Bigram and trigram words are being added to this list. The MPQA subjectivity lexicon includes a word list with sentiment polarity labels. We disseminated word lists from the lexicon that can be used to represent positive, neutral, and negative concepts[16].

III.E Splitting the dataset

Following pre-processing, the tweets are divided into two sets for training and testing. A training set contains 70% of the data, while a testing set has 30%.The next phase is the selection of the unigram, bigram, and trigram features. Using distinct training and testing data sets, we are able to identify unigrams, bigrams, and trigrams.[[links]][.1] Pre-processed data from the data set are then divided into two sub data sets called training and testing data. All of the models employed in this methodology are trained using this training data set, and after the models have classified the data, they are evaluated based on the testing data set. The testing data set evaluates the model based on how well it fits the training data set.

III.F. Training data:

Training data is utilised to fit the sentimental analysis models, and it comprises 75% to 90% of the total data set.

III.G Testing data:

The old data, which ranges from 10% to 35%, is utilised to assess the classification models that were developed using the training data set.

Model building

The phases of model creation that make up the art of a general machine learning sentiment analysis model are as follows. In a typical machine learning model construction process, the training of a machine learning model from raw data to getting predictions from testing data requires the completion of the following phases. The following steps are carried out for each of the machine learning models that have been chosen or chosen for sentiment analysis, including SVM logistic regression, random forest classifier, decision tree classifier, and xgboost classifier.

The phases are

Model fitting

The ML models that we are utilizing in this phase of model construction are really fitted for the training training data set and are afterwards used for performance evaluation.

Prediction

The trend model is tested in this stage of model construction utilizing testing data that is used to evaluate the model's performance.

Training accuracy

Training accuracy is the model's accuracy with respect to the training set of data. Shows how accurate the model is on the training data. The predictions made by the machine learning

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

models are compared to the data set's actual outputs to establish the model's accuracy.

Validation

The process of validation involves confirming that the output in the dataset matches the output anticipated by the model.

Fig 2 Data flow of the system.

Here is one of the machine learning models algorithms along with the other models.

Training Algorithm: Linear SVC

Step 1: Initialize the input data: D = [X, Y], where X is an array of inputs with m features, and Y is an array of corresponding class labels.

Step 2: Set the number of runs for training iterations: num_runs.

Step 3: Set the learning rate: learning_rate = random().

Step 4: Initialize the weight vector: w = zeros(m).

Step 5: Iterate for each run from 1 to num_runs:

- Set the error flag: error = 0.
- Iterate over each data point i from 1 to the length of X:
  - Calculate the predicted class value: pred = dot(X[i], w).
  - If (Y[i] * pred) < 1, then:
    - Update the weight vector: w = w + learning_rate * ((X[i] * Y[i]) - (2 / num_runs) * w).
    - Set the error flag: error = 1.
  - Otherwise:
    - Update the weight vector: w = w + learning_rate * (-2 / num_runs) * w.
- If error is 0 (no errors occurred during this run), then break the loop.

Step 6: Return the fitted model weights, w.

Prediction Algorithm: Linear SVC

Step 1: Take the input data, x, for which the class needs to be predicted.

Step 2: Provide the trained weight vector, w.

Step 3: Calculate the predicted value: pred = dot(x, w).

Step 4: If pred< 0, then assign class1 as the predicted class.

Step 5: Otherwise, assign class2 as the predicted class.

Step 6: Return the predicted class, c.

Evaluation

After applying all machine learning classifiers, including Naive Bayes, Random Forest, Decision Tree, Logistic Regression, Support Vector Classifier, and XGBoost, their performance is assessed through various metrics. Each classifier is evaluated based on Precision, Recall, Accuracy, and F-Measure, using a confusion matrix table. The evaluation is performed using the 10-fold cross-validation method. The entire dataset is divided into 10 equal folds or portions using this technique. Each fold is used as the test data once while the remaining nine folds serve as the training data. This process is repeated 10 times, with each fold being used as the test data exactly once. This ensures that all data points are used for both training and testing. The confusion matrix is used to analyze the performance of the classifiers. It is a table that summarizes the results of the classification task, showing the counts of true positives, true negatives, false positives, and false negatives. The confusion matrix provides a detailed breakdown of the classifier's predictions and the actual class labels. Unfortunately, the details of Figure 2, which explains the contents of the confusion matrix table, are not provided. It would typically include the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. These values are essential for calculating performance metrics such as Precision, Recall, Accuracy, and F-Measure. Overall, the evaluation process involves assessing the performance of each classifier using the 10-fold cross-validation method and analyzing the results using the confusion matrix

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

to gain insights into the classifiers' performance on the given dataset

IV.RESULTS

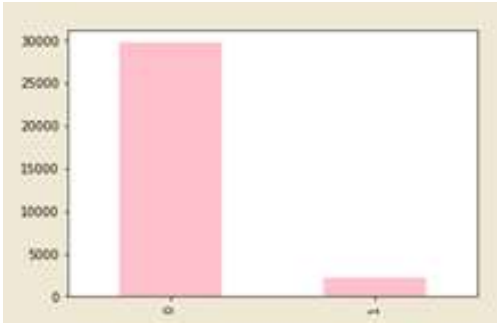

Fig 1 Frequency of occurring words

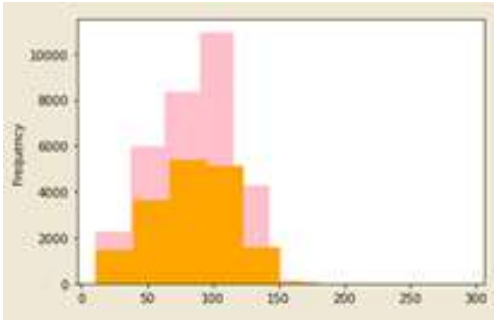

Fig 2 conparition of Postive and Negative tweets

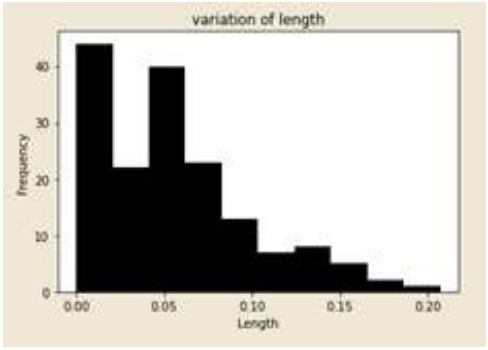

Fig 3 Distribution of tweets.

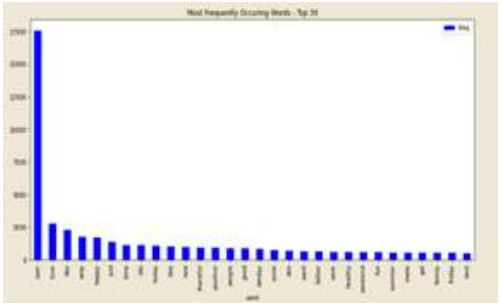

Fig 4 grouping of data based on length

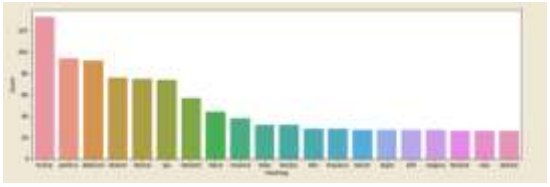

Fig.3.Hastage counting

```
Training Accuracy : 0.9991656585040257
Validation Accuracy : 0.9504442497810036
F1 score : 0.6000000000000001
[[7298  134]
 [ 262  297]]
```

```
Training Accuracy : 0.9851487213716574
Validation Accuracy : 0.9416843949443123
f1 score : 0.5933682373472949
[[7185  247]
 [ 219  340]]
```

```
Training Accuracy : 0.9603687789412206
Validation Accuracy : 0.9555750218996371
f1 score : 0.5748502994011976
[[7396   36]
 [ 319  240]]
```

```
Training Accuracy : 0.978181969880272
Validation Accuracy : 0.9521962207483419
f1 score : 0.4986876640419947
[[7419   13]
 [ 369  190]]
```

Fig.4. performance of the each of the models

V CONCLUSION

The performance of multiple sentiment analysis machine learning models, which were developed using the Twitter dataset, was examined in this research. After cleaning the data set by removing extraneous data from tweets, we preprocessed and chose the necessary features. 90% of the tweets from the Twitter dataset were used to train the sentiment classifier models, and the remaining 10% were utilized to test the trained models. Positive and negative attitudes in tweets are identified via classifiers. The classifiers are based on Machine learning algorithms like SVM classifier, logistic regresion, random forest classifier, Decision tree classifier and XGboost classifier. By using the count vectorization and bagging of words the data is converted into reuired format for the utilization by the models.

V.A.FUTURE ENHANCEMENT

In this study, we conducted a performance analysis of various machine learning models for sentiment analysis using the Twitter dataset. We followed a systematic approach, which included preprocessing the dataset and selecting relevant features to ensure the quality of the data used for training and testing. To begin, we cleaned the Twitter dataset by removing unnecessary

information from the tweets. Next, we split the dataset into a training set (90% of the tweets) and a testing set (10% of the tweets). The training set was used to train the sentiment classifier models, while the testing set was employed to evaluate the performance of the trained models. The classifiers we employed were based on popular machine learning algorithms, such as Support Vector Machine (SVM) classifier, logistic regression, random forest classifier, decision tree classifier, and XGBoost classifier. These classifiers were designed to identify positive and negative sentiments from the tweets. To make the data compatible with the models, we employed techniques like count vectorization and bag-of-words representation. These methods transformed the textual data into a suitable format that could be utilized by the machine learning models. By evaluating the performance of these machine learning models using the provided dataset, we gained insights into their effectiveness in sentiment analysis. The study aimed to identify the strengths and weaknesses of each model, and ultimately provide valuable information for sentiment analysis tasks in the context of Twitter data.

ACKNOWLEDGMENT

We would like to express our gratitude to everyone who assisted us in writing this paper. We would like to thank Lovely Professional University for letting us work on this paper and would like to use this opportunity to thank our mentor for his guidance and support. Last but not least, we would want to express our gratitude to everyone who assisted us in finishing the paper, especially my friends and peers for their unwavering support.

REFERENCES

[1] Pong-Inwong, Chakrit; Songpan, Wararat; (2019) Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining,International journal of machine learning and cybernetics (10):2177-2186

[2] Bing Liu ,"Many Facets of Sentiment Analysis"

[3] Erik Cambria Dipankar Das Sivaji Bandyopadhyay Antonio Feraco;A Practical Guide to Sentiment Analysis

[4] Saif M. Mohammad;"Challenges in Sentiment Analysis"

[5] Rushdi-Saleh, M., Mart´ın-Valdivia, M.T., Urena-L ˜ opez, L.A., Perea-Ortega, J.M., 2011. Oca: Opinion corpus for arabic. Journal of the ´ American Society for Information Science and Technology 62, 2045–2054.

[6]Shoukry, A., Rafea, A., 2012. Sentence-level arabic sentiment analysis, in: Collaboration Technologies and Systems (CTS), 2012 International Conference on, IEEE. pp. 546–550.

[7] Mountassir, A., Benbrahim, H., Berrada, I., 2012. An empirical study to address the problem of unbalanced data sets in sentiment classification, in: Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, IEEE. pp. 3298–3303.

[8] Elawady, R.M., Barakat, S., Elrashidy, N.M., 2014. Different feature selection for sentiment classification. International Journal of Information Science and Intelligent System 3, 137–150.

[9] Omar, N., Albared, M., Al-Shabi, A.Q., Al-Moslmi, T., 2013. Ensemble of classification algorithms for subjectivity and sentiment analysis of arabic customers' reviews. International Journal of Advancements in Computing Technology 5, 77.

[10] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS) 26, 12.

[11] Abdul-Mageed, M., Diab, M.T., Korayem, M., 2011. Subjectivity and sentiment

analysis of modern standard arabic, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics. pp. 587–591.

[12] Badaro, G., Baly, R., Hajj, H., Habash, N., El-Hajj, W., 2014. A large scale arabic sentiment lexicon for arabic opinion mining. ANLP 2014 165.

[13] Eskander, R., Rambow, O., 2015. Slsa: A sentiment lexicon for standard arabic., in: EMNLP, pp. 2545–2550.

[14] Abdul-Mageed, M., Diab, M.T., 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis., in: LREC, pp. 1162–1169.

[15] Ganesh K. Shinde, Vaibhav N. Lokhande, Rasika T. Kalyane, Vikas B. Gore, Umesh M. RautSentiment Analysis on Twitter Hashtag Datasets.

[16] EfthymiosKouloumpis, Theresa Wilson, Johanna Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!”. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media page no. 538-541, 2011