International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Enhanced Assessment of Machine Learning Algorithms for Sentiment Analysis: Investigating Precision, Efficiency, Accuracy, and F-measure

***Tedla Bhavani*[1] *and Sunkaraboina Paramesh*[2]**

[12]Department of Computer Science & Engineering (AI&ML) - CMR Institute of Technology-Hyderabad
E-Mail: tedlabhavani@gmail.com and paramesh5809@gmail.com

ABSTRACT

In this paper, different machine learning-based classification algorithms for sentiment analysis—a task in natural language processing that tries to ascertain the sentiment expressed in a text—are compared. The goal is to evaluate how well various algorithms perform in terms of precision, effectiveness, and generalizability. Various texts from various areas, such as product reviews, social media posts, and customer feedback, make up a benchmark dataset with identified attitudes. The algorithms under investigation are Random Forests, Naive Bayes, Support Vector Machines (SVM), and Neural Networks. The findings demonstrate the distinct advantages and disadvantages of each algorithm. Naive Bayes is suited for large-scale sentiment analysis because of its strong computing efficiency and scalability. SVM operates effectively in cases with non-linear decision boundaries and can handle high-dimensional feature spaces. To increase accuracy and handle noisy data, Random Forests use ensemble learning techniques. Deep learning neural network architectures, in particular, excel at accurately predicting sentiment and capturing intricate linguistic patterns. The examination also emphasizes how the design of features and dataset properties affect algorithm performance. N-grams, word embeddings, and syntactic characteristics all have a big impact on performance. The analysis also takes into account the effects of data preprocessing, class imbalance, and hyperparameter adjustment.

Keywords: Sentiment analysis, machine learning algorithms, comparative analysis, accuracy, efficiency, generalization capability, benchmark dataset.

1. INTRODUCTION

An opinion is a perspective or assessment of a particular matter that has a significant impact on how an individual makes decisions. The "wisdom of crowds" is reflected in the collective opinions, which can serve as a reliable predictor of the future [1]. Opinion is highly valued in many aspects of life since people's beliefs and decisions are always influenced by how other people perceive and assess the world. Sentiment analysis is the practice of identifying opinions or sentiments as good or negative from reviews of products or services provided by businesses, governments, and organizations [2]. Due to its dynamic range of applicability in a variety of sectors, this field has gained a lot of attention from researchers in recent years. The results of sentiment analysis are beneficial in many fields, including marketing, politics, news analytics, etc. [3].

Today's users find it challenging to choose their favorite product because of the wide variety of goods and services available. Product reviews end up being a very helpful resource. Despite people's readiness to express their opinions on the goods, there is still a problem because there are so many reviews overall [3]. This creates a demand for data mining technology to automatically unearth facts and support decision-making. This type of data mining technique uses sentiment analysis to categorize opinions based on the polarity of reviewers' comments [4].

Sentiment analysis, a technique used to categorize sentiments expressed in text, can be approached through two main methods: the lexicon-based approach and the machine learning approach [5]. The lexicon-based approach involves classifying attitudes using training and test datasets. On the other hand, the machine learning approach does not rely on prior training datasets. Instead, it identifies a list of words or phrases that hold semantic values and focuses on identifying patterns in previously unrecognized data [3].

The fact that there are still a lot of difficult and intriguing research problems in this sector makes it more difficult. In contrast to topic-based text classification, sentiment analysis of a document is much more difficult to carry out. Situations often change how one feels and how one expresses their opinions. As a result, a term that is believed to be an opinion may be favorable in one situation but turn negative in another. 'Unpredictable' is a subjective term that has multiple meanings depending on the field. For instance, "an unpredictable plot in the movie" expresses approval of the film, whereas "an unpredictable steering wheel" expresses disapproval of the product, an automobile [3].

Document level, phrase level, and feature level are the three layers of the sentiment categorization process. The entire document is categorized at the level of the document as either positive or negative. Sentence-level sentiment classification takes into account each individual sentence to determine if it is good or negative. The identification and extraction of product attributes from the source data is the focus of feature level sentiment classification

[5].

The objective of this research study is to classify sentiment polarity using three different text classification algorithms: MNB, KNN, and SVM. The classification is based on three sentiment-labeled sentence datasets, namely the Yelp restaurant review dataset, the Amazon cell phones and accessories review dataset, and the IMDB movie review dataset. To enhance processing efficiency and the performance of the sentiment polarity classification algorithms, the data underwent preprocessing steps prior to being used as input. The preprocessing steps included case folding, stop word removal, lemmatization, stemming, feature selection using the chi-squared approach, feature weighting using the TFIDF method, and other necessary preparations. Four performance evaluation metrics were employed to assess the effectiveness of the sentiment polarity categorization algorithms: accuracy, precision, recall, and F-measure. These metrics serve as standards to measure the performance and effectiveness of the algorithms in accurately categorizing sentiment polarity.

2. LITERATURE SURVEY

Similar issues have been studied in the past in many situations, and the area is expanding quickly. In order to achieve the best analysis results to aid in better decision-making, there are many research considerations that must be made. For the purpose of comparing lexical-based and machine learning-based techniques, a novel approach based on SVMs has been developed. This method demonstrates how machine learning-based sentiment classification approaches are highly effective and surpass lexical-based approaches [6].

Both supervised and unsupervised techniques were used to automatically categorize the attitudes of 2000 social network members. The study's findings showed that supervised machine learning methods beat unsupervised machine learning methods with a small classification error [7].

It has been tested to automatically categorize sentiments in text documents using classification techniques. The written papers in this experiment were categorized by topic and overall sentiment into negative and positive attitudes. The results of the researchers' experiment showed that the topic-based sentiment categorization is a difficult task for the classification algorithms [8] [9].

A comparison of approaches for sentiment analysis that rely on rule-based, lexical, and machine learning has been presented. In addition to showing that the cleaner the data, the more accurate the information, this study discovered that machine learning-based approaches outperform both lexical-based and rule-based approaches [10].

In this comparative study, five classification algorithms were evaluated on a movie review dataset to differentiate between genuine and fake reviews. The performance of the algorithms, including SVM, Naive Bayes, KNN, K*, and Decision Tree-J48, was assessed [11]. Among these algorithms, SVM demonstrated superior performance, surpassing the other four methods in accurately distinguishing between authentic and fake reviews.

The sentiment of twitter data can be determined using an approach that blends machine learning-based methods with preprocessing techniques. This method has been presented and its effectiveness against more traditional machine learning techniques has been tested. The proposed approach works better than the customary machine learning approaches, it is discovered [12].

This study incorporates two machine learning algorithms, SVM and NB, and explores the effectiveness of four feature selection strategies: chi-squared, information gain, mutual information, and symmetrical uncertainty. The findings indicate that SVM with the Information Gain strategy outperforms the other methods. Additionally, it is observed that the chi-squared method exhibits better noise tolerance, despite its relative performance compared to the Information Gain strategy [13].

The TF-IDF approach and the KNN algorithm have been proposed as a framework for text classification. This approach yields positive results and offers the opportunity to modernize and enhance the current embedded classification method [14].

Using data from Twitter, it has been suggested that an approach that combines preprocessing with TFIDF weighting can be compared to more traditional machine learning algorithms. The proposed approach surpasses conventional machine learning techniques for sentiment analysis, according to the researchers [15].

3. METHODOLOGY

The polarity of review data can be determined using a variety of techniques. Sentiment analysis based on machine learning is the most popular and efficient method. The polarity of a data piece is determined by the machine learning-based sentiment analysis technique as either a positive class or a negative class. The graphic below shows the actions that were taken during this investigation. They were applied to find the most efficient algorithm and to establish the polarity of the review data.

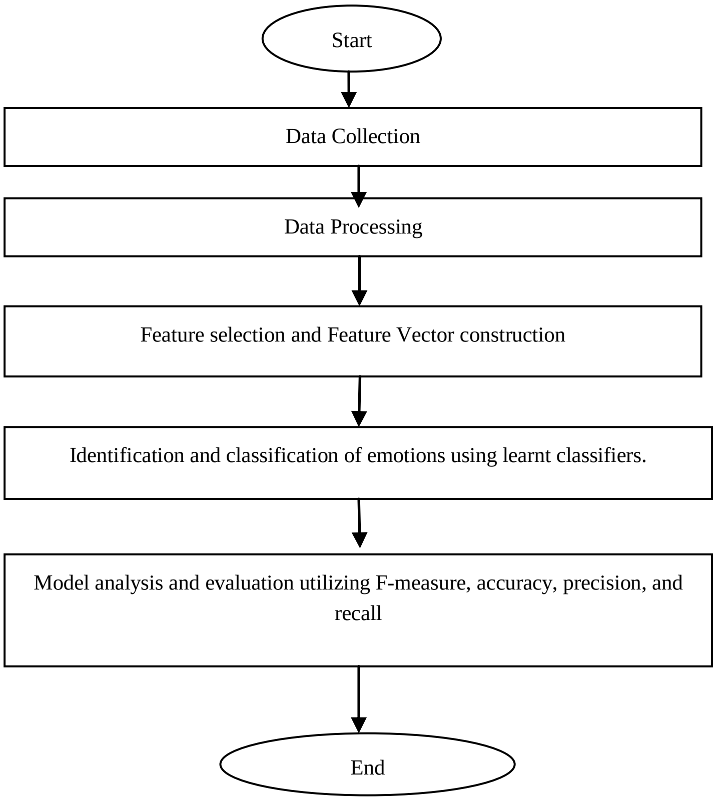
Start
Data Collection
Data Processing
Feature selection and Feature Vector construction
Identification and classification of emotions using learnt classifiers.
Model analysis and evaluation utilizing F-measure, accuracy, precision, and recall
End

Figure 1: Flow of work

Data Collection:

The research utilized three sentiment-labeled sentence datasets, obtained from the Kaggle machine learning repository (kaggle.com). Each dataset varied in size, containing a different number of sentences. While these datasets can be applied to various text classification tasks, the focus of this study was on sentiment analysis. Prior to analysis, the datasets underwent minimal preprocessing, including case folding and stop word removal, to enhance their suitability for the task at hand.

Dataset 1

Data from Yelp restaurant reviews was obtained from the Kaggle machine learning repository and used as the study's initial data set. The data set consists of two characteristics: review and sentiment. The first characteristic measures the value of the review text, while the second measures the value of the sentiment category, with 0 denoting a negative sentiment category and 1 denoting a positive sentiment category for the relevant review text, respectively. This data collection consists of 992 reviews with the associated sentiment categories.

Tools Used

The Pycharm IDE was used to implement each of the three machine learning-based algorithms for sentiment analysis in Python.

Data Preprocessing

The preprocessing phase aims to prepare unstructured text data, specifically reviews, for further processing. The following preprocessing procedures were applied in this study:

Case folding: The text was converted to lowercase to ensure uniformity. For example, "DON'T WASTE YOUR TIME ON THIS 'FILM'" was transformed to "don't waste your time on this 'film'."

Stop word removal: Commonly used words that do not carry significant meaning, such as "don't," "your," "this," and "on," were removed. For instance, "Don't waste your time watching this movie" became "waste time watching movie."

Lemmatization: The words in the text were reduced to their base or root form. For example, "wasting" and "watching" were lemmatized to "waste" and "watch," respectively. The lemmatized sentence would be "do not waste your time watch this movie."

By applying these preprocessing techniques, the text data was standardized and prepared for subsequent analysis and processing.

Feature Vector Construction and Feature Selection

A computer's inability to interpret text data directly is one of its fundamental flaws. Therefore, text data must be represented using numbers. In most cases, terms are utilized as features to denote the text. This causes the text representation to have a high dimension. Features must be filtered to minimize dimension and remove noise in order to increase classification performance and processing effectiveness [16].

TFIDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical weight calculation method used to determine the importance of terms within a set of documents. It combines two metrics: term frequency (TF) and inverse document frequency (IDF). The formula to calculate the TF-IDF weight for a term (ti) in a document (dj) is:

$TF\text{-}IDF(ti, dj) = TF(ti, dj) * IDF(ti)$

Where:

- TF(ti, dj) represents the term frequency of term ti in document dj, which measures how often the term appears in the document.
- IDF(ti) represents the inverse document frequency of term ti, which measures the importance of the term in the entire document set.

The issue with TF-IDF is that it primarily emphasizes the significance of terms in the training set without taking into account their link with particular classes or categories. The chi-square feature selection approach can be used to overcome this restriction.

The chi-square approach aids in determining the degree to which a word (ti) and a class (ck) are related. It determines whether the use of a phrase is affected by the class to which it belongs. The terms that are most closely connected with a given class are determined by computing the chi-square statistic for each term and class.

Chi-Square Method

Chi-square is a method to find the top k features as follows: The formula for the chi-square method, specifically for calculating the chi-square statistic, is as follows:

$\chi^2 = \Sigma [(O - E)^2 / E]$

Where:

- χ^2 represents the chi-square statistic.
- Σ signifies the summation symbol, indicating that the following calculation is performed for each term and class.
- Represents the observed frequency or count of a term occurring in a particular class.
- E represents the expected frequency or count of a term occurring in a particular class, assuming no association between the term and class.

4. RESULTS

In this study, the analysis of the three MLBCAs for sentiment analysis listed above has been compared for the three datasets of sentiment labeled sentences listed above. The comparison is based on four performance criteria: accuracy, precision, recall, and F-measure. After the training phase, all three algorithms were tested on the entire test dataset to obtain the results. The K-parameter for the KNN method has been set at 9 in this research project since it produced better results on this value of K than on other feasible values for the datasets used in the study.

Performance result of MLBCAs for sentiment analysis on dataset1 and their comparison

There are 992 tuples in the dataset1 (i.e., the restaurant review dataset), however after preprocessing only 702 tuples remain, of which 348 fall into the category of one (i.e., positive sentiment) and 354 fall into the category of zero (i.e., negative sentiment). Only 561 tuples were used for training, and the remaining 141 (65 0 and 76 1) were used for testing. The classification report, which was generated after three MLBCAs were applied to the test dataset acquired from dataset1, is displayed in Table 1.

Table 1: Confusion matrix on dataset1

	KNN		MNB		SVM	
	Predicted Positive	Predicted Negative	Predicted Positive	Predicted Negative	Predicted Positive	Predicted Negative
Actual Positive	52	23	59	19	61	17
Actual Negative	17	50	13	54	14	53

The performance results for the three methods applied to dataset 1 were summarized and are presented in Table 2. The table provides an overview of the accuracy, recall, and F-measure values, which are the average metrics calculated from the precision, recall, and F-measure values for each sentiment category, as shown in Table 1 of the article.

Table 2: Performance result on dataset1

Algorithms	Accuracy	Precision	Recall	F-Measure
KNN	74%	74%	74%	74%
MNB	78.7%	79%	78%	77.5%
SVM	78.5%	78.5%	78%	78.5%

Based on Figure 2, the results indicate that SVM achieved a high accuracy rate of 78.5% in sentiment analysis, outperforming KNN, which exhibited a lower accuracy of 74%. SVM also demonstrated high precision and recall levels of 78.5% and 79%, respectively. In contrast, KNN exhibited lower accuracy and recall rates, measuring at 73% for both metrics. Additionally, Figure 2 provides insights into the F-measure as presented in Table 2. Once again, SVM outperformed the other algorithms, achieving an F-measure score of 78.5%, while KNN obtained the lowest score of 74%.

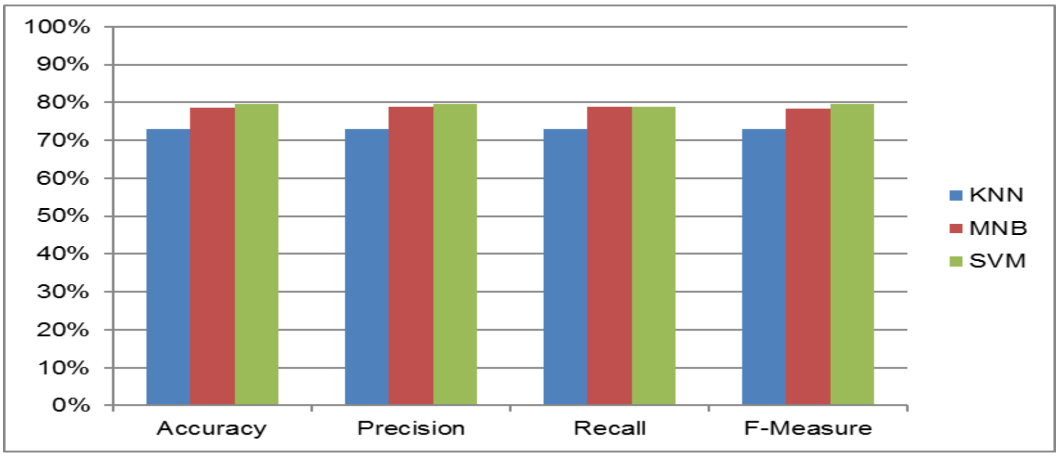

Figure 2. Compare graph of table

5. CONCLUSION

In conclusion, this comparative analysis of machine learning algorithms for sentiment analysis aimed to assess their accuracy, efficiency, and generalization capability. The study utilized a benchmark dataset consisting of diverse texts from various domains, including product reviews, social media posts, and customer feedback. The algorithms investigated were Naive Bayes, Support Vector Machines (SVM), Random Forests, and Neural

Networks. The results of the analysis revealed that each algorithm had its unique strengths and weaknesses. Naive Bayes demonstrated good computational efficiency and scalability, making it suitable for large-scale sentiment analysis tasks. SVM showed robustness in handling high-dimensional feature spaces and performed well when dealing with non-linear decision boundaries. Random Forests leveraged ensemble learning techniques to improve accuracy and effectively handle noisy data. Neural Networks, particularly deep learning architectures, excelled in capturing complex linguistic patterns and achieved high accuracy in sentiment prediction.

REFERENCES

[1] A. Z. H. Khan, A. Mohammad and V. M. Thakare, "Sentiment Analysis Using Support Vector Machine," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 4, pp. 105-108, 2015.

[2] R. CL and G. S, "Machine Learning based Analysis of Twitter Data to Determine a Person's Mental Health Intuitive Wellbeing,"International Journal of Applied Engineering Research, vol. 13, no. 21, pp. 14956-4963, 2018.

[3] M. Ghosh and G. Sanyal, "An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning," Journal of Big Data, vol. 5, 2018.

[4] G. Isabelle, W. Maharani and I. Asror, "Analysis on Opinion Mining Using Combining Lexicon-Based Method and Multinomial Naive Bayes," in International Conference on Industrial Enterprise and System Engineering, 2018.

[5] S. M. Vohra and J. B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques," Journal of Information, Knowledge and Research in Computer Engineering, vol. 2, no. 2, pp. 313-317.

[6] M. Annett and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," in 21st Conference of the Canadian Society for Computational Studies of Intelligence:Advances in Artificial Intelligence, 2008.

[7] L. Lopes, V. Machado and R. Rabelo, "Automatic Cluster Labeling through Artificial Neural Networks," in International Joint Conferenceo on Neural Networks, 2014.

[8] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classifcation using Machine Learning Techniques," in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.

[9] H. P. Rahmath and T. Ahmad, "Sentiment Analysis Techniques - A Comparative Study," International Journal of Computational Engineering & Management, vol. 17, no. 4, pp. 25-29, 2014.

[10] A. Kathuria and S. Upadhyay, "A Novel Review of Various Sentimental Analysis Techniques," International Journal of Computer Science and Mobile Computing, vol. 6, no. 4, pp. 17-22, 2017.

[11] E. Elmurngi and A. Gherbi, "Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques," International Journal on Advances in Systems and Measurements, vol. 11, no. 1 & 2, pp. 196-207, 2018.

[12] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processingMethods on Twitter Sentiment Analysis," IEEE Access, vol. 5, pp. 2870-2879, 2017.

[13] S. D. Sarkar and S. Goswami, "Empirical Study on Filter based Feature Selection Methods for Text Classification," International Journal of Computer Applications, vol. 81, no. 6, pp. 38-43, 2013.

[14] K. Huda, M. T. Nafis and N. K. Shaukat, "Classification Technique for Sentiment Analysis of Twitter Data," International Journal of Advanced Research in Computer Science, vol. 8, no. 5, pp. 2551-2555, 2017.

[15] B. Trstenjaka , S. Mikacb and D. Donkoc, "KNN with TF-IDF Based Framework for Text Categorization," in Procedia Engineering, 2014.

[16] S. Ahmed, S. Hina and R. Asif, "Detection of Sentiment Polarity of Unstructured Multi-Language Text from Social Media, “International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 199-203, 2018.