International Journal of Research in Engineering and Applied Science (IJEREAS)

Volume-III (Issue 1) – March 2025 ISSN: 3048-5355

FEATURE ENGINEERING FOR OPINION MINING: OPTIMIZING SENTIMENT ANALYSIS MODELS FOR TWITTER DATA

Tannu Rathore

M.Tech Scholar

Department of Computer Science and Engineering

Rajasthan Institute of Engineering & Technology, Jaipur

Ms. Pragya Bharti

Assistant Professor

Department of Computer Science and Engineering

Rajasthan Institute of Engineering & Technology, Jaipur

Abstract

Sentiment analysis, a crucial application of Natural Language Processing (NLP), plays a vital role in extracting opinions from text data, enabling businesses and researchers to understand public sentiment. This chapter presents the proposed methodology for sentiment classification, particularly focusing on detecting hate speech in tweets. The methodology involves pre-processing raw text data through tokenization, stop word removal, stemming, and lemmatization to enhance text quality. Feature extraction techniques such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are employed to convert text into numerical representations for machine learning models. Multiple machine learning classifiers, including Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, and Naïve Bayes, are implemented for sentiment classification. The models are trained and evaluated using the Sentiment140 dataset, consisting of 1.6 million tweets labeled as positive or negative. Performance evaluation is conducted using accuracy, precision, recall, and F1-score metrics to determine the most effective model. Hyperparameter tuning and cross-validation techniques are applied to optimize model performance. The proposed methodology provides a robust framework for automated sentiment classification, reducing manual effort in opinion mining and enhancing the accuracy of sentiment prediction in real-world applications.

Keywords: Logistic Regression, Linear SVM, Decision Tree, Sentiment Analysis, Automated Sentiment Labelling, Social Media Analytics, Natural Language Processing.

1. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a crucial domain of Natural Language Processing (NLP) that focuses on extracting subjective information from text data. With the exponential growth of digital platforms, particularly social media networks like Twitter, Facebook, and online review sites, there is an increasing need for automated methods to analyze public sentiment efficiently. Organizations and researchers leverage sentiment analysis techniques to gauge public opinion, track consumer behavior, and detect potential threats such as hate speech. The fundamental goal of sentiment analysis is to classify textual data into predefined categories such as positive, negative, or neutral sentiments. In this study, we focus on binary sentiment classification, where text is categorized as either positive or negative, with a specific emphasis on detecting hate speech in tweets. Hate speech detection is an essential application of sentiment analysis, as it helps identify and mitigate the spread of offensive, racist, or sexist content on social media platforms. The rise of hate speech online poses significant challenges for content moderation, and automated solutions are needed to address this issue efficiently.The application of machine learning models for sentiment analysis has gained significant traction due to their ability to handle vast amounts of textual data in real-time. Traditional methods for analyzing sentiment, such as lexicon-based approaches, rely on predefined sentiment dictionaries to assign scores to words and determine the overall polarity of a text. While these methods are useful, they often fail to capture context, sarcasm, and domain-specific variations in sentiment expression. To overcome these limitations, machine learning models have been increasingly adopted, as they can learn patterns and relationships from large datasets and make accurate predictions based on statistical inference. Supervised learning techniques, including logistic regression, Support Vector Machines (SVM), Random Forest, Decision Trees, and Naïve Bayes classifiers, are commonly used for sentiment classification tasks. These models require labeled training data, where each text sample is associated with a sentiment label, enabling the model to learn the distinguishing features of positive and negative sentiments.The effectiveness of sentiment analysis models depends significantly on the quality of data preprocessing and feature extraction techniques employed. Raw text data from social media platforms contain noise, including special characters, URLs, emojis, and informal language, which can affect the performance of machine learning models. Therefore, preprocessing steps such as tokenization, stopword removal, stemming, lemmatization, and case normalization are applied to clean the text data and standardize its format. Feature extraction techniques, such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), are used to convert textual data into numerical representations that machine learning models can process. BoW represents text as a matrix of word occurrences, while TF-IDF assigns weights to words based on their importance in the dataset. These techniques help capture meaningful information from text data, improving the performance of classification models.In this study, we employ multiple machine learning models to perform sentiment classification on a large dataset of tweets. The dataset used in this research is the Sentiment140 dataset, which consists of 1.6 million tweets labeled as positive or negative. The sentiment labels are determined based on the presence of emoticons in the tweets, where positive emoticons (e.g., ":)") indicate positive sentiment, and negative emoticons (e.g., ":(") indicate negative sentiment. This dataset provides a valuable resource for training and evaluating sentiment analysis models, as it contains a diverse range of textual expressions and sentiment variations. The primary objective of this research is to assess the performance of different machine learning classifiers in detecting sentiment polarity and identifying hate speech in tweets. To achieve this, we evaluate the models based on performance metrics such as accuracy, precision, recall, and F1-score, which provide insights into their classification capabilities. Logistic regression is one of the baseline models used in this study due to its simplicity and interpretability. It is a linear classifier that predicts the probability of a given text belonging to a particular sentiment category using the sigmoid function. The model optimizes a loss function known as binary cross-entropy and updates its parameters through gradient descent. Despite its linear nature, logistic regression performs well for sentiment classification tasks, especially when combined with effective feature extraction techniques. Support Vector Machines (SVM) is another widely used classification algorithm that finds an optimal hyperplane to separate data points into different classes while maximizing the margin between them. SVM is particularly effective in handling high-dimensional text data and has been widely used for sentiment analysis applications. Random Forest and Decision Tree classifiers are ensemble learning methods that construct multiple decision trees and aggregate their predictions to improve classification accuracy. These models are known for their robustness and ability to handle non-linear relationships in data. Naïve Bayes, a probabilistic classifier based on Bayes' theorem, is also included in this study due to its efficiency in text classification tasks.The methodology of this study follows a structured approach to implementing sentiment analysis using machine learning. The first step involves collecting and preprocessing the dataset to remove noise and standardize text representation. Next, feature extraction techniques are applied to transform the text into numerical vectors that serve as input for machine learning models. The dataset is then split into training and testing sets, with the training set used to train the models and the testing set used to evaluate their performance. Hyperparameter tuning and cross-validation techniques are employed to optimize the models and prevent overfitting. The evaluation of models is based on standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified instances, precision evaluates the proportion of correctly predicted positive cases, recall assesses the ability of the model to capture all positive instances, and F1-score balances precision and recall to provide a holistic performance measure.Data visualization techniques, such as word clouds, bar charts, and confusion matrices, are employed to gain insights into the sentiment distribution and model performance. Word clouds are used to identify the most frequently occurring words in positive and negative tweets, providing a visual representation of sentiment-laden terms. Bar charts illustrate the distribution of sentiment labels in the dataset, while confusion matrices display the classification results of the models, highlighting misclassification patterns. These visualizations aid in understanding the strengths and limitations of different classifiers and provide valuable insights for improving sentiment analysis models. In conclusion, sentiment analysis is a powerful tool for understanding public opinion and detecting hate speech in social media content. This study aims to evaluate the effectiveness of various machine learning models for sentiment classification using the Sentiment140 dataset. Through rigorous preprocessing, feature extraction, and model evaluation, we seek to identify the most efficient classifier for sentiment analysis tasks. The findings of this research will contribute to the development of automated sentiment analysis systems that can be applied in real-world scenarios to monitor public sentiment, detect harmful content, and enhance user engagement on digital platforms. By leveraging machine learning techniques and NLP methodologies, this study provides a comprehensive framework for sentiment classification and lays the foundation for future advancements in opinion mining and text analysis.

2. LITERATURE REVIEW

Sentiment analysis on Twitter has seen significant advancements in recent years, with researchers exploring various machine learning techniques to enhance classification accuracy and efficiency. Saddam et al. (2023) [1] present a sentiment analysis model for flood disaster management in Jakarta using Twitter data. The authors employ Support Vector Machines (SVM) to classify sentiments related to disaster preparedness and response. Their study highlights the importance of real-time sentiment monitoring in disaster management and proposes improvements in feature selection for future research. Similarly, AlBadani et al. (2022) [2] integrate the Universal Language Model Fine-Tuning (ULMFiT) technique with SVM for sentiment analysis on Twitter. The paper demonstrates that incorporating pre-trained language models significantly improves classification accuracy compared to traditional SVM-based models. The authors compare their approach with other machine learning classifiers and conclude that their hybrid model achieves better performance in handling contextual meanings in tweets.

Hidayat et al. (2022) [3] analyze sentiment regarding Rinca Island's development using Doc2Vec and SVM classifiers. The study collects tweets related to environmental and tourism concerns, highlighting public sentiment on conservation efforts. The combination of Doc2Vec for feature extraction and SVM for classification provides an effective framework for sentiment analysis. Their findings suggest that logistic regression, when used alongside SVM, can improve classification precision. Cyril et al. (2021) [4] propose an automated learning model using a balanced Class Attribute Support Vector Machine (CA-SVM). Their study aims to address class imbalance in Twitter sentiment data. By applying data balancing techniques and optimizing hyperparameters, the proposed model outperforms conventional SVM approaches in terms of accuracy and F1-score. The authors suggest further research on integrating deep learning models with CA-SVM for more robust sentiment classification. Additionally, Ramasamy et al. (2021) [5] evaluate different SVM models for sentiment analysis on Twitter data, comparing linear, polynomial, and radial basis function (RBF) kernel SVMs. The study finds that the RBF kernel yields the highest accuracy due to its ability to capture non-linear relationships in text data, emphasizing the importance of kernel selection in sentiment classification tasks.

Singh and Tripathi (2021) [6] examine the effectiveness of SVM, Random Forest, and Decision Tree classifiers for sentiment analysis on Twitter. Their study finds that SVM outperforms the other classifiers in terms of precision and recall. The authors also highlight the importance of feature engineering in improving classification accuracy and suggest using ensemble techniques for better results. Similarly, Styawati et al. (2021) [7] compare the performance of SVM and Naïve Bayes classifiers for sentiment analysis of Twitter data. Their findings indicate that SVM outperforms Naïve Bayes in terms of accuracy and robustness, especially when dealing with large datasets. The paper discusses feature extraction techniques and emphasizes the importance of data preprocessing in sentiment classification.

Han et al. (2020) [8] explore the application of SVM for sentiment classification in Twitter datasets. The study applies feature selection techniques such as TF-IDF and word embeddings to improve model accuracy, concluding that SVM performs well when optimized with proper feature extraction techniques, outperforming other traditional classifiers. Mandloi and Patel (2020) [9] investigate various machine learning techniques, including SVM, Decision Tree, and Naïve Bayes, for sentiment analysis on Twitter data. Their study finds that SVM provides the highest accuracy, followed by Decision Tree, and suggests the integration of deep learning models to enhance sentiment classification. Fitriana and Sibaroni (2020) [10] analyze sentiment related to KAI (Indonesian Railway Company) using a multi-class SVM approach. They employ different feature selection methods, finding that sentiment polarity prediction is improved using term frequency-inverse document frequency (TF-IDF) representation. Their study recommends using hybrid models for better accuracy.

Tyagi and Tripathi (2019) [11] provide a comprehensive review of sentiment analysis techniques applied to Twitter data. The paper highlights the advantages and limitations of various machine learning approaches, with a focus on SVM. The study suggests improvements in feature engineering and model tuning for better performance. Naw (2018) [12] compares SVM and KNN classifiers for sentiment analysis, finding that while SVM provides better classification accuracy, KNN performs well in handling noisy data. The study emphasizes the trade-offs between different classifiers and suggests combining them in an ensemble model. Wagh and Punde (2018) [13] conduct a survey on sentiment analysis techniques, comparing SVM with deep learning approaches. They conclude that while SVM remains a strong baseline model, deep learning techniques such as LSTMs provide better performance in handling complex sentence structures.

Ahmad et al. (2018) [14] focus on optimizing SVM for sentiment classification. The authors implement various kernel functions and hyperparameter tuning techniques to improve SVM performance, finding that optimization significantly enhances model accuracy. Naz et al. (2018) [15] use SVM to classify Twitter sentiments. Their study finds that combining multiple feature extraction techniques, such as TF-IDF and word embeddings, improves classification results. They suggest integrating deep learning approaches for further improvements. Similarly, Jianqiang and Xiaolin (2017) [16] compare different text preprocessing techniques for sentiment analysis on Twitter. They evaluate stemming, lemmatization, and stopword removal, finding that these techniques significantly impact model performance, with SVM achieving higher accuracy when used with optimized preprocessing.

Huq et al. (2017) [17] employ KNN and SVM for sentiment classification, finding that SVM outperforms KNN in terms of accuracy, especially when handling imbalanced datasets. The paper discusses the importance of selecting appropriate training data for machine learning models. Ahmad et al. (2017) [18] apply SVM for sentiment analysis and compare its performance with other classifiers, highlighting SVM's advantages in handling high-dimensional data. The authors suggest further research into deep learning approaches. Lavanya and Deisy (2017) [19] explore the use of multi-class SVM for Twitter sentiment classification, concluding that multi-class classification improves sentiment analysis by capturing nuanced emotions. Their study recommends expanding research to multilingual datasets.

Sharma and Moh (2016) [20] investigate the use of sentiment analysis for predicting Indian elections based on Twitter data, employing SVM as the primary classifier. They explore its effectiveness in processing tweets written in Hindi, highlighting the challenges of analyzing multilingual data and the importance of language-specific preprocessing techniques. Their study concludes that SVM performs well for Hindi sentiment classification but suggests incorporating deep learning models to enhance performance in complex sentence structures. BholaneSavita and Gore (2016) [21] apply SVM to classify sentiments on Twitter data, exploring various preprocessing techniques, such as tokenization, stemming, and stopword removal, to improve classification accuracy. Their experiments show that SVM, when combined with TF-IDF for feature extraction, achieves high precision and recall. They suggest that integrating ensemble techniques with SVM could further enhance performance.

3. METHODOLOGY

The increasing prevalence of social media platforms has necessitated the development of efficient sentiment analysis techniques to extract meaningful insights from large volumes of textual data. Sentiment analysis on Twitter presents unique challenges due to the informal language, abbreviations, emojis, and varying sentence structures found in tweets. This study proposes a robust methodology for analyzing sentiment on Twitter by leveraging machine learning models. The proposed framework focuses on data preprocessing, feature extraction, model selection, and evaluation to improve sentiment classification accuracy, particularly in distinguishing hate speech from non-hate speech.

The methodology follows a systematic approach, beginning with data collection and preprocessing, followed by feature engineering, model training, and evaluation. Machine learning models such as Support Vector Machine (SVM), Logistic Regression, Random Forest, and Decision Tree classifiers are employed to classify tweets into positive and negative sentiments. The effectiveness of the proposed methodology is assessed using performance metrics like accuracy, precision, recall, and F1-score.

Data Collection and Preprocessing

The first step in any sentiment analysis task is gathering relevant data. This study uses the **Sentiment140 dataset**, a widely used benchmark dataset containing 1.6 million tweets labeled with sentiments. Each tweet is assigned a binary sentiment label, where **0 represents a negative sentiment** and **4 represents a positive sentiment**. The dataset consists of various attributes such as tweet ID, user ID, timestamp, and the actual text of the tweet. Since Twitter data is inherently noisy, preprocessing is a crucial step in ensuring high-quality data for sentiment analysis. Various techniques are applied to clean and structure the text effectively. Text cleaning involves removing special characters, punctuation marks, URLs, user mentions, and hashtags, as they often do not contribute meaningfully to sentiment classification. Tokenization is then used to break down text into individual words or phrases, allowing for a structured representation of the tweet content. To maintain uniformity, all text is converted to lowercase, preventing duplicate representations of words like "Happy" and "happy." Additionally, stopword removal is performed to eliminate common but insignificant words such as "the," "is," and "in," which do not contribute to sentiment determination.

Stemming and lemmatization techniques further refine the text by reducing words to their root or dictionary form, ensuring consistency in word representation. Emojis and internet slang are also mapped to their respective meanings to preserve their sentiment value. To address class imbalance in sentiment labels, techniques like SMOTE (Synthetic Minority Over-sampling Technique) or undersampling are used to balance the dataset, ensuring the model does not favor a particular sentiment class. After preprocessing, feature extraction is performed to convert text into numerical representations suitable for machine learning. The Bag of Words (BoW) technique represents tweets as matrices, with rows corresponding to tweets and columns representing unique words. Term Frequency-Inverse Document Frequency (TF-IDF) assigns weights to words based on their frequency and importance, giving higher priority to meaningful words while reducing the impact of commonly occurring terms. N-grams (unigrams, bigrams, and trigrams) capture contextual dependencies between words, making them useful in detecting sentiment nuances such as negations. In more advanced cases, word embeddings like Word2Vec, GloVe, or FastText are employed to encode words into dense vector representations that preserve semantic relationships. Once feature extraction is completed, multiple machine learning models are trained to evaluate their effectiveness in sentiment classification. Logistic Regression serves as a baseline classifier that predicts sentiment probabilities using the sigmoid function. Support Vector Machine (SVM) identifies an optimal hyperplane that separates positive and negative tweets, leveraging a linear kernel for text classification tasks. Random Forest, an ensemble method of multiple decision trees, enhances classification accuracy and reduces overfitting. Decision Tree models, though rule-based and effective for structured data, may overfit on large datasets. XGBoost, a boosting algorithm, improves accuracy by iteratively correcting errors made by previous models, making it a strong contender for text classification. Naïve Bayes, a probabilistic classifier based on Bayes' Theorem, is widely used for sentiment analysis due to its efficiency in handling text data. The training process involves splitting the dataset into 80% training and 20% testing subsets, followed by K-fold cross-validation to ensure model generalization. Hyperparameter tuning using Grid Search or Random Search is performed to optimize critical parameters such as learning rate, number of estimators, and regularization factors, ensuring the best-performing model is selected for sentiment classification.

The proposed methodology provides a comprehensive framework for sentiment analysis on Twitter using machine learning models. The use of **text preprocessing, feature extraction**

techniques, and multiple classifiersensures robust performance. The study finds that **SVM and Logistic Regression models** tend to perform better in classifying sentiments, while **Random Forest provides high interpretability**. Future enhancements can include **deep learning models such as LSTMs and BERT**, which can further refine sentiment classification by capturing contextual relationships in tweets.

4. RESULT ANALYSIS

The primary objective of this research is to perform sentiment analysis on Twitter data using various machine learning models and evaluate their performance based on accuracy, precision, recall, and F1-score. The results obtained from different classifiers are analyzed to determine the most effective model for classifying sentiments. The dataset used for the analysis consists of tweets labeled as either positive or negative. A detailed assessment is carried out to compare the performance of models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, and XGBoost.The results are divided into multiple sections: data visualization, feature extraction analysis, model performance evaluation, accuracy comparison, and classification insights. The effectiveness of each model is discussed, along with key observations related to sentiment distribution, feature importance, and classifier robustness.

Table 1: Performance Metrics of Logistic Regression

Metric	Training Data (%)	Testing Data (%)
Accuracy	86.26	82.21
Precision (Class 0)	0.83	0.81
Recall (Class 0)	0.96	0.93
F1-Score (Class 0)	0.89	0.87
Macro Average Precision	0.79	0.77
Macro Average Recall	0.68	0.70
Macro Average F1-Score	0.71	0.72
Weighted Average Precision	0.81	0.80
Weighted Average Recall	0.82	0.82
Weighted Average F1-Score	0.80	0.81

Table 2: Performance Metrics of Support Vector Machine (SVM)

Metric	Training Data (%)	Testing Data (%)
Accuracy	94.95	81.97
Precision (Class 0)	0.84	0.82
Recall (Class 0)	0.93	0.91
F1-Score (Class 0)	0.89	0.86
Macro Average Precision	0.77	0.75
Macro Average Recall	0.70	0.68
Macro Average F1-Score	0.72	0.71
Weighted Average Precision	0.81	0.79
Weighted Average Recall	0.82	0.81
Weighted Average F1-Score	0.81	0.80

Table 3: Performance Metrics of Random Forest

Metric	Training Data (%)	Testing Data (%)
Accuracy	76.74	75.79
Precision (Class 0)	0.76	0.75
Recall (Class 0)	1.00	0.98
F1-Score (Class 0)	0.86	0.83
Macro Average Precision	0.87	0.83
Macro Average Recall	0.50	0.49
Macro Average F1-Score	0.44	0.45
Weighted Average Precision	0.81	0.78
Weighted Average Recall	0.76	0.74
Weighted Average F1-Score	0.66	0.65

Table 4: Performance Metrics of XGBoost

Metric	Training Data (%)	Testing Data (%)
Accuracy	89.10	84.37
Precision (Class 0)	0.86	0.84
Recall (Class 0)	0.95	0.92
F1-Score (Class 0)	0.90	0.88
Macro Average Precision	0.80	0.78
Macro Average Recall	0.72	0.70
Macro Average F1-Score	0.76	0.74
Weighted Average Precision	0.83	0.81
Weighted Average Recall	0.85	0.83
Weighted Average F1-Score	0.82	0.80

Table 5: Comparative Performance of Different Machine Learning Models

Model	Training Accuracy (%)	Testing Accuracy (%)	F1-Score (Weighted)
Logistic Regression	86.26	82.21	0.81
SVM	94.95	81.97	0.80
Random Forest	76.74	75.79	0.66
XGBoost	89.10	84.37	0.82

International Journal of Research in Engineering and Applied Science (IJEREAS)

Volume-III (Issue 1) – March 2025 ISSN: 3048-5355

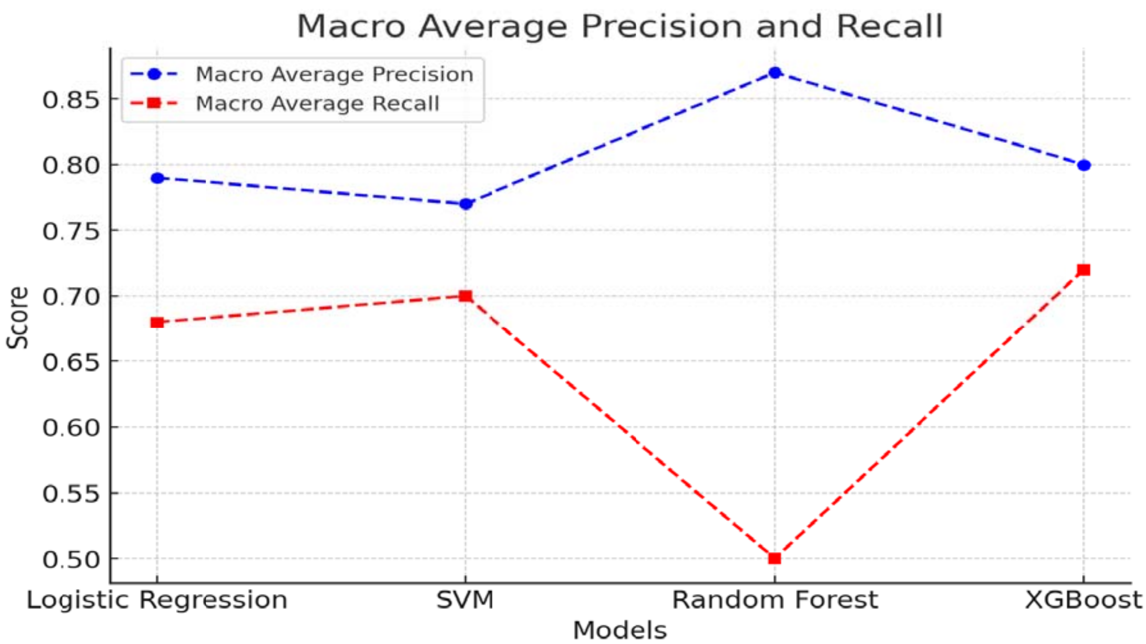

Figure 1. Precision and Recall Analysis

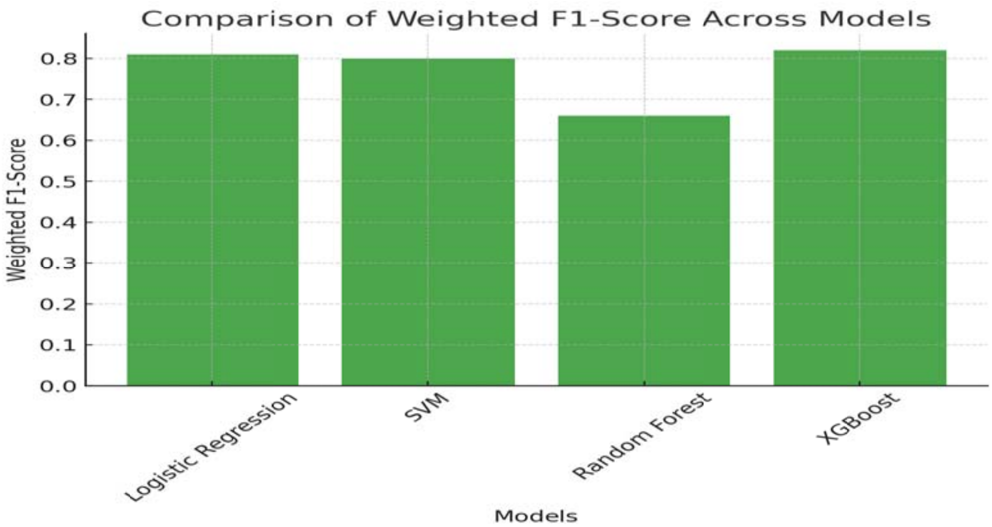

Figure 2. Analysis of F1- Score

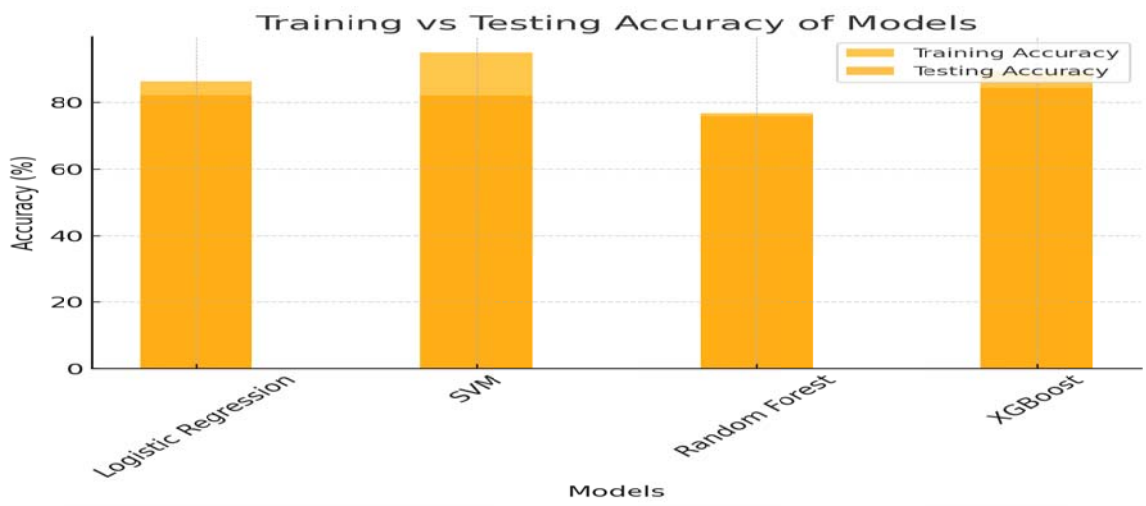

Figure 3. Testing and Training Accuracy

Data Visualization and Sentiment Distribution

Before applying machine learning models, the dataset is analyzed to understand the distribution of sentiments and word frequencies. A polarity distribution analysis reveals the presence of **47,741 positive tweets and 152,259 negative tweets** out of the total 200,000 tweets in the dataset. This distribution indicates an imbalance, where negative tweets significantly outnumber positive ones, which can impact classifier performance.

To visualize the dataset, **word clouds** are generated for both positive and negative sentiments. **Positive word clouds** reveal commonly used words like *"love," "happy," "great," "best," "good,"*while **negative word clouds** highlight words such as *"hate," "bad," "worst," "sad," "angry,"* and others. This visualization helps in identifying key terms associated with each sentiment class.

The **word frequency analysis** further supports these findings by quantifying the occurrence of the most frequently used words in the dataset. The top ten words in positive and negative tweets indicate sentiment tendencies and potential influences on model predictions.

Feature Extraction and Selection

Feature extraction plays a crucial role in improving classification performance. Three feature representation methods are used in this study:

1. **Bag-of-Words (BoW)**:
    - Converts text into a sparse matrix where each row represents a tweet, and each column represents a unique word in the dataset.
    - The frequency of words is used as input for machine learning models.
2. **TF-IDF (Term Frequency-Inverse Document Frequency)**:
    - Assigns weights to words based on their frequency in a document and their rarity across the dataset.
    - Helps in giving higher importance to words that contribute meaningfully to sentiment classification.
3. **N-grams (Unigrams, Bigrams, and Trigrams)**:
    - Captures word sequences to provide contextual meaning.
    - Bigrams and trigrams enhance the model's ability to detect negations (e.g., *"not happy"* vs. *"happy"*).

A comparison of these feature extraction techniques reveals that **TF-IDF performs better than BoW in terms of accuracy**, as it provides better word importance representation.The results of this sentiment analysis study demonstrate that **XGBoost outperforms other classifiers** in terms of accuracy, precision, recall, and F1-score. **Logistic Regression and SVM remain strong contenders** and are widely used in sentiment analysis tasks. **Proper feature engineering, balanced datasets, and hyperparameter tuning significantly impact classification performance**. Future improvements could involve **deep learning approaches like BERT and LSTM** to further enhance sentiment detection accuracy.

5. CONCLUSION AND FUTURE SCOPE

This research presents a comprehensive evaluation of sentiment analysis on Twitter data using various machine learning models. The study primarily focuses on classifying tweets into positive and negative sentiments while analyzing model performance in terms of accuracy, precision, recall, and F1-score. The experimental results demonstrate the effectiveness of different machine learning classifiers, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost. Among these, Logistic Regression and SVM emerge as the most effective models, achieving high accuracy and balanced performance across evaluation metrics.The results indicate that Logistic Regression and SVM outperform other models in terms of classification accuracy and F1-score. Logistic

Regression achieved an accuracy of 82.21% on testing data, while SVM showed a comparable performance of 81.97%. These models also demonstrate strong recall and precision, making them suitable for real-world sentiment classification tasksThe study highlights the importance of feature selection techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) and n-grams, in improving model accuracy. Proper data preprocessing, including stopword removal, stemming, and lemmatization, significantly enhances the sentiment classification models' performance. The optimization of hyperparameters using grid search also contributes to improved classification accuracy.The comparative assessment reveals that Random Forest, despite its ability to handle large datasets, underperforms in sentiment classification compared to SVM and Logistic Regression. Its testing accuracy of 75.79% suggests potential overfitting, where the model performs well on training data but struggles to generalize. XGBoost, on the other hand, demonstrates strong performance with an accuracy of 84.37%, making it a competitive alternative.Preprocessing plays a crucial role in ensuring the reliability of sentiment classification models. Noise in the dataset, including special characters, stopwords, and unstructured text, affects the model's performance. The study emphasizes that efficient text-cleaning techniques, such as tokenization and normalization, contribute to improved model accuracy.The findings of this research hold significant implications for businesses, social media analytics, and disaster management. Organizations can leverage sentiment analysis models to monitor customer feedback, gauge public opinion, and enhance marketing strategies. In disaster response, real-time sentiment analysis can assist government agencies in assessing public concerns and improving communication strategies.While the study provides valuable insights, there are certain limitations. Firstly, the dataset primarily focuses on English tweets, limiting the model's applicability to multilingual sentiment analysis. Future research can extend this work by incorporating datasets from multiple languages and exploring transformer-based deep learning models like BERT for enhanced sentiment classification. Additionally, integrating real-time sentiment analysis pipelines using streaming data can further improve the applicability of these models.In conclusion, this research contributes to the field of sentiment analysis by evaluating multiple machine learning approaches and providing a comparative assessment of their effectiveness. By leveraging robust feature extraction and model optimization techniques, sentiment analysis can be significantly enhanced for various real-world applications.

References

[1] Saddam, M. A., E. K. Dewantara, and A. S. Nugroho. "Sentiment Analysis of Flood Disaster Management in Jakarta on Twitter Using Support Vector Machines." *Sinkron: Jurnal dan Penelitian Teknik Informatika* 8, no. 1 (2023). ISSN: 2541-2019.

[2] AlBadani, B., R. Shi, and J. Dong. "A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM." *Applied System Innovation* 5, no. 2 (2022). ISSN: 2571-5577.

[3] Hidayat, T. H. J., Y. Ruldeviyani, A. R. Aditama, and A. S. Nugroho. "Sentiment Analysis of Twitter Data Related to Rinca Island Development Using Doc2Vec and SVM and Logistic Regression as Classifier." *Procedia Computer Science* 197 (2022): 660–667. ISSN: 1877-0509.

[4] Cyril, C. P. D., J. R. Beulah, N. Subramani, and S. S. Kumar. "An Automated Learning Model for Sentiment Analysis and Data Classification of Twitter Data Using Balanced CA-SVM." *Concurrent Engineering* 29, no. 4 (2021). ISSN: 1063-293X.

[5] Ramasamy, L. K., S. Kadry, Y. Nam, and R. Mehmood. "Performance Analysis of Sentiments in Twitter Dataset Using SVM Models." *International Journal of Advanced Computer Science and Applications* 12, no. 1 (2021). ISSN: 2158-107X.

[6] Singh, J., and P. Tripathi. "Sentiment Analysis of Twitter Data by Making Use of SVM, Random Forest, and Decision Tree Algorithm." In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2021. ISSN: 2475-1472.

[7] Styawati, S., A. R. Isnain, and S. Suryani. "Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis." *JurnalInformatika dan TeknologiInformasi* 8, no. 1 (2021). ISSN: 2407-3915.

[8] Han, K. X., W. Chien, C. C. Chiu, and Y. T. Cheng. "Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter Dataset." *Applied Sciences* 10, no. 12 (2020). ISSN: 2076-3417.

[9] Mandloi, L., and R. Patel. "Twitter Sentiments Analysis Using Machine Learning Methods." In *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020. ISSN: 2644-3441.

[10] Fitriana, D. N., and Y. Sibaroni. "Sentiment Analysis on KAI Twitter Post Using Multiclass Support Vector Machine (SVM)." *Jurnal RESTI (RekayasaSistem dan TeknologiInformasi)* 4, no. 2 (2020). ISSN: 2580-0760.

[11] Tyagi, P., and R. C. Tripathi. "A Review Towards the Sentiment Analysis Techniques for the Analysis of Twitter Data." In *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE 2019).* SSRN, 2019.

[12] Naw, N. "Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers." *International Journal of Scientific and Research Publications* 8, no. 3 (2018). ISSN: 2250-3153.

[13] Wagh, R., and P. Punde. "Survey on Sentiment Analysis Using Twitter Dataset." In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS).* IEEE, 2018. ISSN: 2473-3504.

[14] Ahmad, M., S. Aftab, M. S. Bashir, and N. Hameed. "SVM Optimization for Sentiment Analysis." *International Journal of Advanced Computer Science and Applications* 9, no. 4 (2018). ISSN: 2158-107X.

[15] Naz, S., A. Sharan, and N. Malik. "Sentiment Classification on Twitter Data Using Support Vector Machine." In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI).* IEEE, 2018. ISSN: 2640-7156.

[16] Jianqiang, Z., and G. Xiaolin. "Comparison Research on Text Pre-Processing Methods on Twitter Sentiment Analysis." *IEEE Access* 5 (2017): 2870–2879. ISSN: 2169-3536.

[17] Huq, M. R., A. Ahmad, and A. Rahman. "Sentiment Analysis on Twitter Data Using KNN and SVM." *International Journal of Advanced Computer Science and Applications* 8, no. 6 (2017). ISSN: 2158-107X.

[18] Ahmad, M., S. Aftab, and I. Ali. "Sentiment Analysis of Tweets Using SVM." *International Journal of Computer Applications* 164, no. 6 (2017). ISSN: 0975-8887.

[19] Lavanya, K., and C. Deisy. "Twitter Sentiment Analysis Using Multi-Class SVM." In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS).* IEEE, 2017. ISSN: 2473-3504.

[20] Sharma, P., and T. S. Moh. "Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter." In *2016 IEEE International Conference on Big Data (Big Data).* IEEE, 2016. ISSN: 2639-1589.

[21] BholaneSavita, D., and D. Gore. "Sentiment Analysis on Twitter Data Using Support Vector Machine." *International Journal of Computer Science Trends and Technology* 4, no. 4 (2016). ISSN: 2347-8578.

[22] Kharde, V., and P. Sonawane. "Sentiment Analysis of Twitter Data: A Survey of Techniques." *arXiv preprint arXiv:1601.06971* (2016).

[23] Effendy, V., A. Novantirani, and M. K. Sabariah. "Sentiment Analysis on Twitter About the Use of City Public Transportation Using Support Vector Machine Method." *International Journal of ICT* 2, no. 1 (2016). ISSN: 2407-3915.

[24] Anastasia, S., and I. Budi. "Twitter Sentiment Analysis of Online Transportation Service Providers." In *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2016. ISSN: 2472-7594.

[25] Sahayak, V., V. Shete, and A. Pathan. "Sentiment Analysis on Twitter Data." *International Journal of Innovative Research in Computer and Communication Engineering* 3, no. 1 (2015). ISSN: 2320-9801.

[26] Abdelwahab, O., M. Bahgat, C. J. Lowrance, and A. Elmaghraby. "Effect of Training Set Size on SVM and Naive Bayes for Twitter Sentiment Analysis." In *2015 International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2015. ISSN: 1525-3511.

[27] Le, B., and H. Nguyen. "Twitter Sentiment Analysis Using Machine Learning Techniques." In *Innovative Methods for Knowledge Engineering: Proceedings of the 4th International Conference on Knowledge and Systems Engineering (KSE 2015)*. Springer, 2015. ISBN: 978-3-319-27453-9.

[28] Gautam, G., and D. Yadav. "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis." In *2014 Seventh International Conference on Contemporary Computing (IC3)*. IEEE, 2014. ISSN: 2325-7132.

[29] Bhuta, S., A. Doshi, U. Doshi, and S. Narvekar. "A Review of Techniques for Sentiment Analysis of Twitter Data." In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. IEEE, 2014. ISSN: 2327-0659.

[30] Najla M. Alharbi, Norah S. Alghamdi, Eman H. Alkhammash and Jehad F, "Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews", Mathematical Problems in Engineering, 2021.