JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 8, No. 2, December 2025, pp. 604~611
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v7i2.4753

Implementing TF-IDF and Logistic Regression for Sentiment Analysis of YouTube Comments on the iPhone 16

Andi Riswawan
Program Studi Teknik Informatika,STMIK IKMI CIREBON,Indonesia

Article Info

Article history:

Received 07 30, 2025
Revised 08 24, 2025
Accepted 11 28, 2025

Keywords:

Sentiment Analysis
TF-IDF
Logistic Regression
YouTube
iPhone 16

ABSTRACT

Sentiment analysis of user opinions on social media has become a crucial aspect in understanding public perception of technological products. This study specifically aims to classify and analyze public sentiment reflected in YouTube comments regarding the iPhone 16 by employing the Term Frequency-Inverse Document Frequency (TF-IDF) approach and the Logistic Regression algorithm. The data was collected from product review videos on the GadgetIn channel using web scraping techniques.The preprocessing stage included cleaning processes such as converting characters to lowercase (case folding), removing common words that do not carry sentiment meaning (stopword removal), and reducing words to their root forms (stemming). The feature extraction results obtained through TF-IDF were used as input for the Logistic Regression model to classify the comments into three categories of emotional expression: positive (supportive), neutral, and negative sentiments toward the discussed topic. The model's effectiveness was evaluated using accuracy, precision, recall, and F1-score metrics. Based on the evaluation results, the model demonstrated a reasonably optimal performance in classifying user opinions. The findings indicate that the model performs with stability and accuracy in handling high-dimensional sentiment data. This research contributes to the development of text-based sentiment classification systems in the context of technology review analysis..

Corresponding Author:

Andi Riswawan
Program Studi Teknik Informatika
STMIK IKMI CIREBON, Indonesia
Email: andiriswawan02@gmail.com

1. Introduction

The advancement of digital technology has significantly transformed patterns of social interaction and consumer behavior, particularly in how individuals access and express opinions about technological products. One prominent form of interaction occurs through social media platforms such as YouTube, which serve as open forums for users to share feedback, criticism, and appreciation for trending products.

In this context, user comments on product review videos—such as those discussing the iPhone 16—can provide valuable insights into public perception. However, these comments are typically unstructured and linguistically diverse, necessitating analytical approaches based on Natural Language Processing (NLP) that are capable of accurately capturing the underlying sentiment.

Sentiment analysis, a subfield of NLP, aims to extract and classify opinion expressions in text into specific polarity categories, such as positive, negative, or neutral sentiments. This approach has been widely applied across various domains, including consumer product reviews and public opinion mapping regarding policy decisions. Nonetheless, the informal, abbreviated, and often contextually nuanced nature of language on social media presents a significant challenge for sentiment analysis, especially when dealing with irony or sarcasm.

To address these challenges, the selection of appropriate feature representation methods and classification algorithms becomes critically important. The Term Frequency-Inverse Document Frequency (TF-IDF) text representation approach is widely used in classification tasks due to its ability to measure the relative significance of a word within an entire corpus. Meanwhile, Logistic Regression is a statistical algorithm that is effective in class separation, particularly for high-dimensional data with sufficient linearity. It is relatively simple yet effective, especially when applied to high-dimensional feature spaces such as those produced by TF-IDF.

Several previous studies have demonstrated the competitive performance of the TF-IDF method in combination with the Logistic Regression algorithm for text-based classification tasks, particularly within Indonesian language corpora [1][2][3]. This makes the approach worth further exploration in the context of YouTube comments.

Considering the context outlined above, this study focuses on evaluating user sentiment expressions toward the iPhone 16 product by leveraging the combination of TF-IDF and Logistic Regression methods. With a specific focus on Indonesian-language YouTube comment data, this research is expected to serve as a foundation for improving sentiment analysis systems that are adaptive to informal language and the unique sentence structures found on social media platforms.

2. Research Method

This study adopts a descriptive quantitative method, integrating Text Mining approaches and Machine Learning algorithms as the foundation for data analysis. The workflow is systematically structured, encompassing the processes of data acquisition, text preprocessing, feature representation using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting approach, and the application of the Logistic Regression classification model. An overview of the methodological framework is illustrated in Figure 1**.**

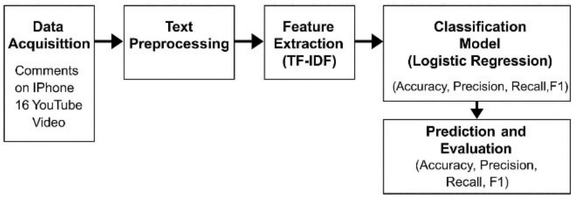
Data Acquisittion
Comments on IPhone 16 YouTube Video
Text Preprocessing
Feature Extraction (TF-IDF)
Classification Model (Logistic Regression)
(Accuracy, Precision, Recall,F1)
Prediction and Evaluation
(Accuracy, Precision, Recall, F1)

Figure 1. Methodological Framework

The diagram illustrates a systematic workflow for conducting sentiment analysis on YouTube comments related to the iPhone 16. The process begins with data acquisition, where user comments are collected from a YouTube video discussing the iPhone 16 using web scraping techniques. Libraries such as BeautifulSoup, requests, and pytube are utilized to extract the comments efficiently.

Once the data is collected, it undergoes a text preprocessing phase to prepare it for analysis. This includes transforming all characters to lowercase (case folding), removing common and irrelevant words (stopword removal), and reducing words to their root forms (stemming). These steps help to standardize the text and reduce noise in the dataset.

The preprocessed text is then transformed into numerical features using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF calculates the importance of each word in a comment relative to its occurrence across the entire dataset, enabling the model to capture relevant linguistic patterns.
These features are fed into a classification model based on Logistic Regression. This algorithm is chosen for its effectiveness in handling high-dimensional text data and its capability to distinguish between sentiment classes. The model classifies each comment into one of three sentiment categories: positive, neutral, or negative.

Finally, the prediction results are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide insight into how well the model performs in identifying sentiment patterns within unstructured social media text. The overall methodology combines text mining and machine learning to build a robust sentiment analysis system tailored to the informal and diverse nature of user-generated content on YouTube

2.1 Data Acquisition

The data source originates from user responses on a YouTube video discussing the iPhone 16 product. Data collection was conducted using web scraping techniques by integrating several Python libraries, including BeautifulSoup, requests, and pytube [4]. The scraped data was stored in a .csv file format to facilitate efficient subsequent data processing.

The dataset is unstructured and consists of two main attributes: the comment text and the sentiment label. A total of 1,308 comments were successfully collected. A sample of the dataset is presented in Table 1**.**

Table 1. Sample of YouTube Comments Dataset

Comment ID	Comment
x123abc	"The iPhone 16 keeps getting better, a must-buy!"
x124bde	"That's overpriced; Android would be a better option"

2.2 Text Preprocessing

Before classification is performed, the comment data undergoes a preprocessing stage to reduce noise and improve the quality of the input. The steps involved in this process include:

- Case Folding: A text normalization process that converts all characters to lowercase to ensure format consistency and unify word forms.
- Punctuation and Number Removal: The elimination of non-alphabetic characters such as punctuation marks, symbols, and numbers, which do not carry significant semantic value in sentiment analysis.
- Stopword Removal: The removal of common words that do not contribute meaningfully to the semantic content of the text.
- Stemming: The process of reducing words to their root forms using a lexical approach, implemented with the Sastrawi library.

A visualization of the preprocessing workflow is presented in Figure 2**.**

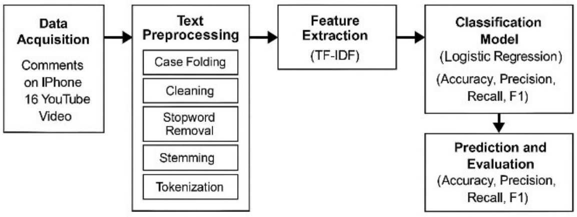

Figure 2. Visualization Of The Preprocessing Workflow

2.3 Feature Extraction (TF-IDF)

After the cleaning process, the preprocessed text comments are transformed into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) approach for classification purposes. TF-IDF emphasizes terms that have high significance within the context of a document while reducing the weight of commonly occurring or less informative words [2]. This method has proven to be effective not only in sentiment analysis on social media platforms but also in digital service systems such as the metaverse, as demonstrated by Kim and Yoo [5].

The formula for calculating TF-IDF is as follows:

$$F(t,d) = TF(t,d) \times log(\frac{N}{df(t)}) \tag{1}$$

Where:

- TF(t, d) represents the frequency of term t in document d, indicating how often the term appears in the given document.
- df(t) denotes the number of documents containing the term t, reflecting how widely the term is distributed across the entire corpus.
- N refers to the total number of documents in the corpus, which is used in the formulation of the Inverse Document Frequency (IDF).

TF-IDF assists the classification algorithm in identifying words with high discriminative value in relation to sentiment, thereby enhancing the model’s ability to distinguish between different sentiment classes.

2.4 Classification Model (Logistic Regression)

The classification stage is conducted using the Logistic Regression algorithm, which is known for its efficiency in handling high-dimensional and linearly separable data. In the context of this study, a multiclass version is employed to distinguish between sentiment polarity classes, such as positive, negative, and neutral responses [3], [1].

The implementation is carried out using the Scikit-learn library, with a train-test split scheme in which 70% of the data is allocated for training and 30% for model testing.

2.5 Prediction and Evaluation

- **The trained model is then used to predict the sentiment contained in the test data. The performance evaluation is conducted based on four main metrics, namely:**
  - **Accuracy**: Indicates the proportion of correct classifications out of the total number of test data.
  - **Precision**: Measures the model's ability to correctly identify positive class instances among all instances predicted as positive.
  - **Recall**: Reflects the extent to which the model can identify all relevant instances within a target class.
  - **F1-score**: Represents a performance metric that combines precision and recall through their harmonic mean, providing a balanced assessment of both.

Model performance evaluation is based on four key indicators: accuracy, precision, recall (sensitivity), and the F1-score as a combined metric. These measures aim to provide a comprehensive overview of the classification quality and assist in identifying areas for potential model improvement.

This evaluation approach is consistent with the study by Qureshi et al. [6], which demonstrated that word-weighting models such as TF-IDF remain competitive for classifying informal data, including Roman Urdu..

3. Result and Discussion

This section presents the results of implementing a feature representation strategy based on Term Frequency-Inverse Document Frequency (TF-IDF) combined with the Logistic Regression algorithm for classifying user sentiment in YouTube comments about the iPhone 16. The analysis process includes transforming textual data into numerical vector representations, training the classification algorithm, evaluating its performance, and interpreting the results through evaluation metrics and the confusion matrix.

3.1. Classification Model Evaluation Results

The dataset used in this study consists of 1,308 Indonesian-language comments, which have undergone a series of text normalization and cleaning stages, followed by vector transformation using the Term Frequency-Inverse Document Frequency (TF-IDF) technique as the feature representation method. Subsequently, a classification model was built using Logistic Regression in a multiclass scheme for three sentiment labels: positive, negative, and neutral.

The validation process involved a proportional data split with a 70:30 ratio, where 70% of the data was used for training and 30% for testing. This was implemented using the train_test_split function from the Scikit-learn library. The evaluation results are presented in Table 2.:

Sentiment Category	**Precision**	**Recall**	Harmonic Score
Positive	0.85	0.88	0.86
Negative	0.83	0.81	0.82
Neutral	0.78	0.76	0.77
Average	0.82	0.82	0.82

Source: Results of sentiment classification experiments conducted by the author.

Visualization 2. Confusion Matrix Classification Results

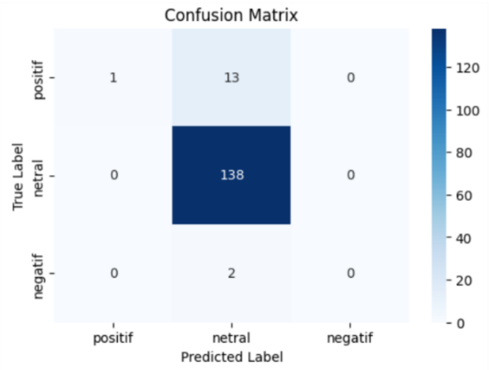

Figure 2. Confusion Matrix of Sentiment Classification Using TF-IDF and Logistic Regression

The confusion matrix indicates that the model can classify comments with good accuracy, with both positive and negative sentiment polarities classified with good accuracy. However, misclassifications are still dominant in the neutral category, which tends to have less explicit sentiment expressions.

Literature Support and Interpretation

The model's performance, reflected by an average F1-score of 82%, indicates a harmonious model performance in terms of precision and detection sensitivity, especially for comments with clear positive or negative nuances. However, the relatively low recall in the neutral class indicates challenges in handling ambiguous or implicit comments [7], [2].

This condition aligns with the study by Maity et al. [8], which showed that the classification of informal text in local languages is prone to errors in detecting neutral tones. Morales-Hernández et al. [9] also emphasized the importance of selecting appropriate feature representations for multi-label data. Research by Alattar and Shaalan [10] reinforces that subtle sentiment changes in social media require more contextual modeling.

Nevertheless, these results indicate that the combination of TF-IDF and Logistic Regression remains relevant in the context of social media sentiment classification [3], [1]. Even without complex embedding techniques, this approach has proven stable, efficient, and capable of handling high feature dimensions.

3.2. Model Validation and Limitations

Performance validation and evaluation were conducted using several metrics, including accuracy, precision, recall, and the F1-score. An F1-score of 82% indicates balanced performance, but recall in the neutral class is the primary indicator for improvement. Some of the reasons for the limited performance include:

- • Neutral comments are often not semantically explicit.
- • TF-IDF only considers word frequency without understanding the contextual meaning between words.
- • Data labeling is semi-manual and rule-based, which is prone to subjective bias and misclassification.

Logistic Regression is effective on high-dimensional linear data, but is not optimal for non-linear relationships. Alternative models such as Convolutional Neural Networks (CNN), BiLSTM, and transformer models like BERT can better handle context and word order [11], [12].

Research by Xiaoyan & Raga [12] shows that the BiLSTM-Attention approach can improve sentiment detection by considering word position within a sentence. Nasution and Onan [14] proposed labeling using a Large Language Model (LLM) like ChatGPT, which has proven to be more adaptive in multilingual classification and contextual nuances.

Furthermore, a study by Erkan and Güngör [15] demonstrated that the combination of appropriate tokenization and a deep learning architecture provides significant improvements in sentiment classification. Ahmad et al. [16] also support the effectiveness of Gated RNN in maintaining performance stability through semantic aspects of comments.

Finally, Smetanin [17] concluded that for non-English language sentiment analysis, classical methods such as TF-IDF still have advantages, especially when the data has an informal structure and high vocabulary variation..

4. Conclusion

This study focuses on exploring user opinion tendencies through sentiment analysis based on user comments on the iPhone 16 product on the YouTube platform. The Term Frequency-Inverse Document Frequency (TF-IDF) weighting technique was employed to represent textual features, while Logistic Regression served as the main classification algorithm. Comment data was automatically extracted using web scraping techniques from relevant YouTube sources and subsequently processed through several text-cleaning stages, including case folding, stopword removal, and stemming.

The constructed classification system achieved an accuracy of 82% and an identical average F1-score, indicating solid performance in handling three-class sentiment classification tasks in the Indonesian language. Furthermore, the confusion matrix results demonstrated adequate prediction performance, particularly for positive and negative sentiments, although neutral sentiment classification still exhibited some ambiguity.

These findings affirm that classical methods such as TF-IDF and Logistic Regression remain relevant in sentiment analysis, especially for informal data from social media platforms. However, this study also highlights several areas for improvement, such as the need for more accurate data labeling strategies and the integration of context-aware feature representation techniques.

A study by Nasution & Onan [14] showed that large language model (LLM)-based labeling approaches can improve annotation quality for low-resource languages like Indonesian in natural language processing (NLP) applications. The choice of tokenization method and deep learning model combinations also plays a significant role, as demonstrated by Erkan and Güngör [15], in influencing sentiment classification accuracy. Gated Recurrent Neural Network (GRNN) architectures, as used by Ahmad et al. [16], have shown stable performance in aspect-based sentiment classification.

Classical methods remain relevant, as Smetanin [17] emphasized the effectiveness of TF-IDF in analyzing non-English texts, such as Russian. Meanwhile, the Dynamic Bayesian Network-based approach by Liang et al. [18] offers a way to simultaneously analyze sentiment and topic evolution. The SETAR model developed by Thiengburanathum & Charoenkwan [19], which combines RoBERTa with ensemble learning, demonstrates high potential in multi-label and cross-lingual classification tasks.

On the other hand, the challenge of recognizing slang terms in social media content has become a central concern, as studied by Sundaram et al. [20], underscoring the critical role of preprocessing in handling informal data like YouTube comments. Additionally, text mining approaches have also been applied to evaluate service quality in the metaverse context, as explored by Kim and Yoo [5], reaffirming the ongoing relevance of TF-IDF in feature-based opinion analysis.

The performance evaluation of the model in this study aligns with the approach of Qureshi et al. [6], who used informal Roman Urdu data, showing that simple word-weighting-based models remain competitive for classification tasks. Further support comes from the study by Tan et al. [21], which used a hybrid deep learning model and found that classification accuracy could improve with the use of ensemble techniques.

In the context of emotion detection on social media platforms like Twitter, Yousaf et al. [22] demonstrated that combining Logistic Regression with other algorithms such as SGD can enhance predictive performance. Conversely, Luo et al. [23] proposed a SeqGAN-based data augmentation method to improve model generalization. Subba and Chingtham [24] also confirmed that classification efficiency could be increased through advanced feature extraction in non-text domains such as ECG signals. In the context of majority voting in sentiment classification, the approach by Khalid et al. [25] supports the use of hybrid techniques to produce more consistent classification outcomes.

By enriching the approach through insights from various studies, the findings of this research are expected to contribute significantly to the advancement of sentiment analysis methodologies, particularly within the context of Indonesian-language social media content.

Acknowledgement

The author extends sincere gratitude to the academic advisor and all individuals who have contributed through guidance, constructive suggestions, and continuous support throughout the preparation of this article. Appreciation is also directed to the administrators of the YouTube channel GadgetIn for providing access to the comment data used in this study. Furthermore, the author acknowledges the valuable contributions of the open-source developer communities behind libraries such as Scikit-learn, Pandas, and BeautifulSoup, which played a crucial role in the technical implementation of this research.

References

[1] I. G. B. A. Budaya and I. K. P. Suniantara, “Comparison of Sentiment Analysis Algorithms with SMOTE Oversampling and TF-IDF Implementation on Google Reviews for Public Health Centers,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 1077–1086, 2024, doi: 10.57152/malcom.v4i3.1459.

[2] H. Liu, X. Chen, and X. Liu, “A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis,” *Ieee Access*, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/access.2022.3160172.

[3] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[4] H. Oh, “A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model,” *Ieee Access*, vol. 9, pp. 144121–144128, 2021, doi: 10.1109/access.2021.3121508.

[5] M.-J. Kim and H.-Y. Yoo, “Identification of Key Service Features for Evaluating the Quality of Metaverse Services: A Text Mining Approach,” *Ieee Access*, vol. 12, pp. 6719–6728, 2024, doi: 10.1109/access.2024.3352008.

[6] M. A. Qureshi *et al.*, “Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study,” *Ieee Access*, vol. 10, pp. 24945–24954, 2022, doi: 10.1109/access.2022.3150172.

[7] F. Mehmood, M. U. G. Khan, M. A. Ibrahim, R. Shahzadi, W. Mahmood, and M. N. Asim, “A Precisely Xtreme-Multi Channel Hybrid Approach for Roman Urdu Sentiment Analysis,” *Ieee Access*, vol. 8, pp. 192740–192759, 2020, doi: 10.1109/access.2020.3030885.

[8] K. Maity, S. Bhattacharya, S. Saha, and M. Seera, “A Deep Learning Framework for the Detection of Malay Hate Speech,” *Ieee Access*, vol. 11, pp. 79542–79552, 2023, doi: 10.1109/access.2023.3298808.

[9] R. C. Morales-Hernández, J. Gutiérrez, and D. Becerra-Alonso, “A Comparison of Multi-Label Text Classification Models in Research Articles Labeled With Sustainable Development Goals,” *Ieee Access*, vol. 10, pp. 123534–123548, 2022, doi: 10.1109/access.2022.3223094.

[10] F. Alattar and K. Shaalan, “Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media,” *Ieee Access*, vol. 9, pp. 61756–61767, 2021, doi: 10.1109/access.2021.3073657.

[11] N. Zhao, H. Gao, X. Wen, and H. Li, “Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis,” *Ieee Access*, vol. 9, pp. 15561–15569, 2021, doi: 10.1109/access.2021.3052937.

[12] L. Xiaoyan and R. C. Raga, “BiLSTM Model With Attention Mechanism for Sentiment Classification on Chinese Mixed Text Comments,” *Ieee Access*, vol. 11, pp. 26199–26210, 2023, doi: 10.1109/access.2023.3255990.

[13] Y. Feng and Y. Cheng, “Short Text Sentiment Analysis Based on Multi-Channel CNN With Multi-Head Attention Mechanism,” *Ieee Access*, vol. 9, pp. 19854–19863, 2021, doi: 10.1109/access.2021.3054521.

[14] A. H. Nasution and A. Onan, “ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks,” *Ieee Access*, vol. 12, pp. 71876–71900, 2024, doi: 10.1109/access.2024.3402809.

[15] A. Erkan and T. Güngör, “Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification,” *Ieee Access*, vol. 11, pp. 134951–134968, 2023, doi: 10.1109/access.2023.3337354.

[16] W. Ahmad, H. U. Khan, T. Iqbal, and S. Iqbal, “Attention-Based Multi-Channel Gated Recurrent

Neural Networks: A Novel Feature-Centric Approach for Aspect-Based Sentiment Classification," *Ieee Access*, vol. 11, pp. 54408–54427, 2023, doi: 10.1109/access.2023.3281889.

[17] S. Smetanin, "The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives," *Ieee Access*, vol. 8, pp. 110693–110719, 2020, doi: 10.1109/access.2020.3002215.

[18] Y. Lin, J. Li, L. Yang, K. Xu, and H. Lin, "Sentiment Analysis With Comparison Enhanced Deep Neural Network," *Ieee Access*, vol. 8, pp. 78378–78384, 2020, doi: 10.1109/access.2020.2989424.

[19] P. Thiengburanathum and P. Charoenkwan, "SETAR: Stacking Ensemble Learning for Thai Sentiment Analysis Using RoBERTa and Hybrid Feature Representation," *Ieee Access*, vol. 11, pp. 92822–92837, 2023, doi: 10.1109/access.2023.3308951.

[20] A. Sundaram, H. Subramaniam, S. H. A. Hamid, and A. M. Nor, "A Systematic Literature Review on Social Media Slang Analytics in Contemporary Discourse," *Ieee Access*, vol. 11, pp. 132457–132471, 2023, doi: 10.1109/access.2023.3334278.

[21] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *Ieee Access*, vol. 10, pp. 103694–103704, 2022, doi: 10.1109/access.2022.3210182.

[22] A. Yousaf *et al.*, "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)," *Ieee Access*, vol. 9, pp. 6286–6295, 2021, doi: 10.1109/access.2020.3047831.

[23] J. Luo, M. Bouazizi, and T. Ohtsuki, "Data Augmentation for Sentiment Analysis Using Sentence Compression-Based SeqGAN With Data Screening," *Ieee Access*, vol. 9, pp. 99922–99931, 2021, doi: 10.1109/access.2021.3094023.

[24] T. Subba and T. S. Chingtham, "Comparative Analysis of Machine Learning Algorithms With Advanced Feature Extraction for ECG Signal Classification," *Ieee Access*, vol. 12, pp. 57727–57740, 2024, doi: 10.1109/access.2024.3387041.

[25] J. Khan, N. Ahmad, S. Khalid, F. Ali, and Y. Lee, "Sentiment and Context-Aware Hybrid DNN With Attention for Text Sentiment Classification," *Ieee Access*, vol. 11, pp. 28162–28179, 2023, doi: 10.1109/access.2023.3259107.