JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 02
2025

Research Article
https://doi.org/10.56979/902/2025

Echoes of Opinion: Leveraging Advanced Deep Learning for Spotify Review Classification

Afifa Hameed[1*], Ali Saeed[1], Amina Yaseen[2], and Beenish Zafar[3]

[1]Department of Software Engineering, Faculty of Information and Technology, University of Central Punjab, Lahore, 54000, Pakistan.
[2]Department of Software Engineering Wing, Punjab IT Board, Arfa Kareem Tower, Lahore, 54000, Pakistan.
[3]Department of Computer Science, Faculty of Information and Technology, University of Central Punjab, Lahore, 54000, Pakistan.
[*]Corresponding Author: Afifa Hameed. Email: afifa.hameed@ucp.edu.pk

Received: June 19, 2025 **Accepted:** August 27, 2025

Abstract: The escalation of user-generated content on various streaming platforms, such as Spotify, has created invaluable insights into users' feedback to improve user engagement and interaction. The proposed study aims to classify the Spotify user reviews and subsequently perform sentiment analysis by gathering a dataset from Kaggle containing the users' reviews and appropriate labels. Initially, feature extraction techniques such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Part of Speech Tagging (POS) have been applied. Later, several Machine Learning, Deep Learning, and Transformer-based models have been deployed to classify the reviews. The proposed approach achieved 90 % accuracy with Support Vector Machine (SVM), while among deep learning models, Long Short Terms Memory (LSTM) outperformed Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU) by obtaining 95 % accuracy score. Bert large, on the other hand, surpassed Bert-uncased with 94 % accuracy. These valuable findings of the proposed approach enhance the understanding of feedback analysis for streaming services and further recommend future direction to improve Spotify's recommendation system.

Keywords: Spotify Review; Classification; Machine Learning, Natural Language Processing, Deep Learning, Transformer-Based Model

1. Introduction

The music streaming industry has achieved considerable transformation in recent years, with different platforms such as Spotify that endeavor to transform the way users discover, listen to, and engage their selves in their favorite music, having around 600 million active users [1]. Spotify has been offering a wide collection of songs that, in particular, involves personalized playlists, and has generated music enthusiasts across the world. In this context, the **Figure 1** demonstrates the radical increase in the number of monthly active users of Spotify over the past decade, i.e. from 2015 to 2024 [2]. Regardless of this global approval, considering the feedback of users on a bigger scale remains a key challenge in the domain of Natural Language Processing (NLP). To review the textual data manually becomes time-consuming and ineffective; thus, the proposed approach focuses on the automation of sentiment classification using advanced machine learning as well as deep learning approaches.

Machine learning (ML), conversely, has become a crucial tool in almost every other domain, particularly NLP, due to its prognostic nature, i.e., learning from data and making the relevant predictions [3]. The automation has risen unbelievably over the decade in multiple disciplines, mainly in health care [4], the business industry [5], applications of industrial [6], cybersecurity [7], and various other domains. ML, thus, has publicized auspicious results specifically in the domain of NLP when the task of classifying textual data, sentiment analysis, auto-summarization, and predicting the text based on the input given by

the user, is done through automation [8]. Similarly, deploying ML algorithms to accurately classify the reviews of users on the Spotify app into positive or negative classes will aid in the recognition of the trend as well as the sentiments of the user, therefore, revealing the underlying factors that affect user satisfaction and engagement with the Spotify platform. Alongside ML, Deep Learning (DL) and Transformer based model have also shown promising results over the past few years in the domain of NLP. The main aim of the proposed study is: 1) to assess the conventional approaches for sentiment classification of Spotify reviews; 2) to collate the feature extraction methods such as BoW, TF-IDF, and PoS tagging; and 3) to determine the model that balances the performance with computational efficacy.

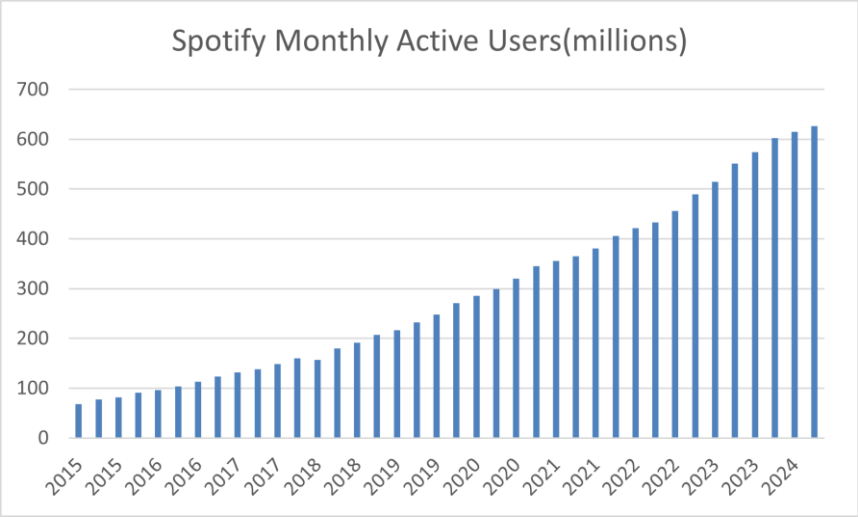


Figure 1. Active Users of Spotify

The proposed research highlights the importance of applying different ML models, such as SVM, NB, RF, KNN, DT, LR, and Deep Learning (DL) algorithms, such as RNN, LSTM, and GRU, to evaluate the performance and efficacy of these models, and resultantly, the study has demonstrated favorable results for review classification [9]. By leveraging the dataset of user reviews of Spotify from Kaggle, the goal is to identify the most effective approach for sentiment analysis [10]. Additionally, the proposed study aims to discover the association between the user sentiments and the key features that the platform provides, such as playlist personalization, the interface of the platform, and the possible ways to discover music. The correlation will aid in getting valuable insights, which will help future researchers and practitioners to make suitable decisions for improving the platform.

The novelty of the proposed study has been highlighted as:

- Features have been extracted from the dataset using different feature extraction techniques, such as TF-IDF, BOW, and PoS Tagging, required mainly for ML algorithms.
- Later, Sentiment analysis has been conducted by exploiting several machine-learning models, namely LR, RF, SVM, KNN, DT, and NB.
- Numerous Deep learning models and their variants have been deployed to perform the task of sentiment analysis of Spotify reviews, namely RNN, LSTM, and GRU.
- Moreover, the transformer-based models such as BERT and BERT-large have also been applied to perform sentiment analysis.

The remainder of the paper is organized as follows: Section Related Work reviews the existing related work for the Spotify recommendation system. The subsequent section provides the details of the dataset used and the proposed methodology exploited in the presented study. Later, the Results and Discussion section outlines an outcome after deploying the models along with the detailed discussion, whereas the Section Conclusion and Future Work finally conclude the paper, and provide future implications in the respective domain.

2. Related Work

Spotify has transmuted the way people adapt themselves to music by offering millions of tracks and songs just a click away. Users often admire its intuitive interface and the personalized recommendations enabled by applying advanced ML and DL algorithms. One of the prominent features of Spotify, presented

as the "Discover Weekly" playlist, has become an outstanding feature among the listeners by tailoring the fresh tracks based on the user's preferences and taste, every Monday [6]. Furthermore, Spotify's ability to create a collaborative playlist for its users to integrate social sharing has nurtured a sense of community among music listeners. Both the free and premium versions and tiers ensure accessibility for a broader audience, although the free version does come with a lot of commercials and interruptions that some users might find uncomfortable and disruptive [7]. Nonetheless, as it can be perceived from the user reviews that it faces some challenges [11], many users criticize that the platform isn't good enough for its audio quality on the free tier, which is lower compared to competitors [10]. Furthermore, several users have experienced frustration with certain issues related to app glitches, or that the customization options are very limited with respect to playlists [8]. Notwithstanding the loopholes, the platform is still considered a favorite platform for music lovers [2]. Inclusive, Spotify has recognized itself as a front-runner, having a massive library as well as pioneering features [12]. Although it faces opposition from many other platforms, such as Amazon, Apple Music, and Deezer but its robust emphasis on user experience, along with music discovery, maintains a lead [13] [14].

Even though various researchers have endeavored to categorize user feedback among different classes on the Spotify platform, using traditional machine learning as well as deep learning methods, the outcome achieved has continued to recede compared to other state-of-the-art methods in domains other than NLP [15]. For example, a study investigated two popular machine learning algorithms, Naïve Bayes and K-Nearest Neighbor, applied to reviews of the Spotify app, achieving an accuracy rate of 81% [16]. At the same time, another study emphasized the shortcomings of the proposed models in capturing the contextual subtleties of user sentiment [12]. Similarly, another study proposed a TF-IDF-based classification framework for Spotify podcasts using traditional ML-based techniques, but eventually reached poor accuracy levels, which shows the need for more complex representation learning [9].

These results imply that while early attempts have shown that sentiment analysis for Spotify data is feasible, the models frequently fall short of generalization because of the lack of contextual embeddings and the restricted diversity of features [17]. To get beyond these restrictions and produce more reliable sentiment categorization, the new study builds on earlier research by combining sophisticated deep learning with transformer-based architectures, like LSTM and BERT. XLM-RoBERTa demonstrated the best overall performance, while Distil BERT achieved the maximum accuracy in this study's sentiment analysis of Spotify app reviews using transformer-based models and natural language processing [18]. One more direction that users often highlight in their reviews is that Spotify's podcast integration has become a crucial feature in recent years [19] [20]**.** Another study examined the effects of architectural design on accuracy, efficiency, and sequential data processing in text-based tasks by comparing 1D and 2D CNN architectures for sentiment analysis of Spotify reviews [21]. While discussing the exploitation of transformer-based models recently by several researchers, another study explores the usage of BERT models for sentiment analysis, showcasing their robust performance and enhanced accuracy following refinement. It also highlights the potential avenues for further research and real-world applications [22].

Although previous researchers have widely discovered the classification of sentiment analysis in domains such as e-commerce and social media [23], inadequate consideration has been given to feedback provided by users on music streaming platforms using unified multi-model frameworks. The current state-of-the-art methods often emphasize on either conventional ML or DL independently, without a reasonable assessment across model families. By comparing conventional neural networks and transformer-based architectures under unvarying feature extraction conditions, the proposed study fills these existing gaps and provides a comprehensive standard for textual sentiment analysis in the streaming space.

3. Materials and Methods

For the proposed methodology, various algorithms have been deployed on the dataset of Spotify in order to perform app review classification.

3.1. Dataset Description

The dataset utilized in the proposed study contains the user reviews regarding the famous Spotify musical platform, and the dataset was collected from Kaggle. The dataset contains a total of 51,473 rows, each representing a unique user review for the Spotify application. The dataset has two columns: Review: This column contains the text of user reviews, reflecting their experiences, opinions, and feedback on the

Spotify app, and Sentiment Label that categorizes each review as either "positive" or "negative" based on its sentiment.

The sentiment distribution in the dataset is such that negative sentiments are 56 % of the total reviews, while positive sentiments are 44 % of the total reviews. The sample of the Spotify dataset is demonstrated in **Table 1**.

Table 1. Sample of Spotify Dataset

Sr. No	Review	Label
1	Great music service, the audio is high quality.	Positive
2	Please ignore the previous negative rating. This is really a nice app.	Positive
3	This pop-up Get the best Spotify experience of the app" is really annoying	Negative
4	Really buggy and difficult to use in the recently updated version	Negative
5	Dear Spotify, why do I get songs that I didn't want to listen to?	Negative

Furthermore, the distribution of the number of reviews with respect to positive and negative sentiments is demonstrated in **Figure 2**.

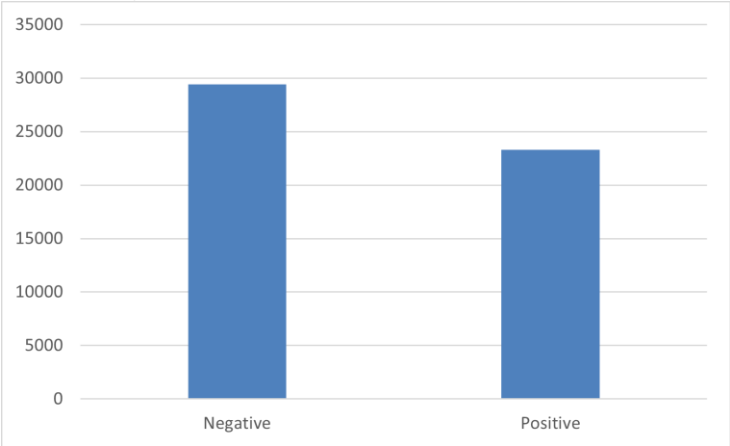


Figure 2. Sentiment Count for Spotify Dataset

3.2. Feature Engineering

Feature Engineering plays an integral part in the performance of the proposed model, as different feature engineering techniques may yield different results. In this test, three types of feature engineering techniques were exploited: Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), & Part-of-Speech (POS) tagging. The BoW way turns word reviews into number lists. It determines the frequency of the word, as to how often each word shows up. This strategy helps the models spot shared trends as well as feelings. TF-IDF further fine-tunes this by giving weight to the words. It identifies the importance of each word in the relevant context. POS tagging, on the other hand, adds grammar data by spotting the role of each word. The POS technique further helps the models in comprehending the syntactic relationships as well as structural relationships among the words. The combination of these feature engineering techniques produces a robust set of features. This, as a result, significantly enhances the performance evaluation of the proposed models, such as SVM, RF, and NB, while ensuring more relevant identification of positive, negative, and neutral comments within the Spotify reviews.

Extensive preprocessing was done to improve the extracted features' quality. To reduce noise and guarantee data consistency, all text was converted to lowercase and stop words, punctuation, and special characters were eliminated. By taking these actions, the features were able to lessen the impact of superfluous variants while precisely capturing the text's semantic meaning. The feature set successfully captured both lexical and syntactic properties by integrating these preprocessing techniques with Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and Part-of-Speech (POS) tagging. The models' robustness and overall classification performance were enhanced by this integration, which allowed them to identify minute changes in sentiment.

Experimental results across several classifiers further confirmed the efficacy of the feature engineering process. Support Vector Machines (SVM), which obtained the highest accuracy and F_1 scores among the models evaluated, continuously outperformed the others, especially when BoW and POS features were used. With TF-IDF features improving their precision and recall in identifying positive and negative feelings, Random Forest and Naive Bayes also demonstrated competitive performance. These results demonstrate how important it is to apply a variety of complementary feature engineering techniques to textual datasets since they offer richer representations of user opinions, which makes sentiment classification in the context of Spotify reviews more accurate and dependable.

3.3. Proposed Methodology

The initial step of the proposed methodology involves acquiring data, which was sourced from the Spotify music app through Kaggle. This dataset was subsequently preprocessed to make it appropriate for analysis. The preprocessing stage consists of several sub-steps, including data cleaning and normalization, as well as refining the dataset to eliminate outliers, noise, and any missing information, ensuring high-quality data is supplied to the model for improved accuracy. Following this, various feature extraction methods were employed, such as TF-IDF, Bag of Words, and Part of Speech tagging. These methods were instrumental in capturing the linguistic features of the textual data, facilitating the creation of feature vectors that are crucial for model training and evaluation. The resulting feature vector serves as a robust foundation for implementing a range of advanced models.

As a result, the models that have been applied in the proposed methodology are divided mainly into three categories: machine learning models, deep learning models, and transformer-based models. The models that outperformed the other conventional models are Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and the transformer-based BERT model. SVMs are recognized for their resilience in managing high-dimensional feature spaces produced by BoW and Term Frequency-TF-IDF. Conversely, LSTM networks effectively capture sequential dependencies and contextual dynamics within textual data, thereby overcoming the shortcomings of conventional frequency-based methodologies. Accuracy, precision, recall, and F_1-score are among the common evaluation metrics used to gauge each model's performance once it has been trained and evaluated using the generated feature representations. The trade-offs between sophisticated deep learning architectures and conventional machine learning techniques are discussed in this comparative analysis, which guarantees both interpretability and predictive accuracy. Figure 3 visualizes the flow of the proposed methodology.

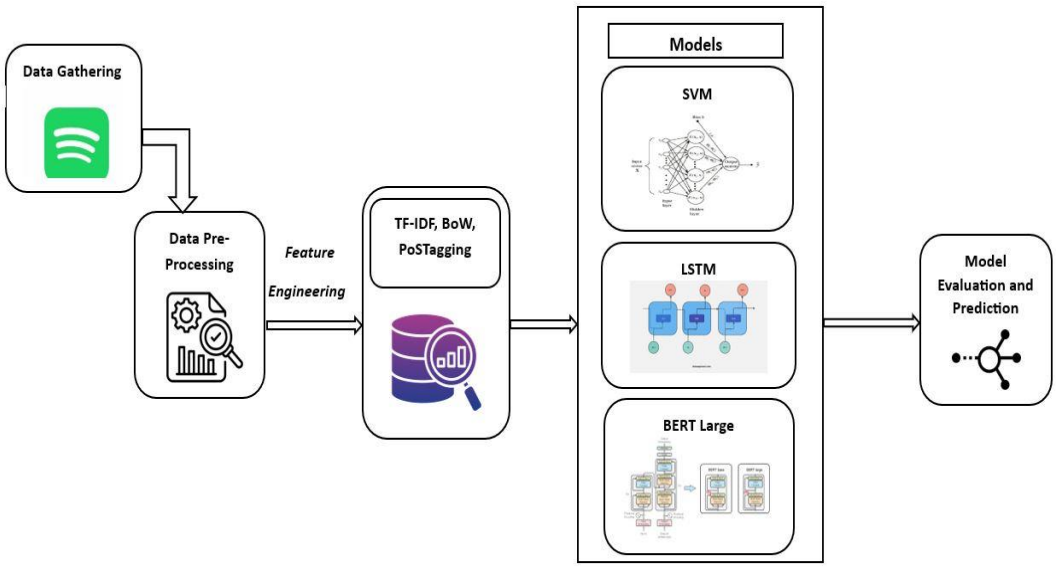


Figure 3. Proposed Methodology

3.4. Machine Learning Algorithms

The following section elaborates on the list of six widely known machine learning algorithms that have been applied to our dataset in order to evaluate the effectiveness of all the state-of-the-art models.

3.4.1. Random Forest (RF)

Random Forest, a widely recognized machine learning algorithm, belongs to the class of supervised learning techniques. It is employed for both classification and regression tasks. In classification, Random Forest processes categorical data, while for regression, it deals with continuous data. RF is based on a vast number of decision trees, where each decision tree gives a one-class prediction, and that class needs to have the maximum number of predictions in order for the model to predict accurately [24]. The algorithm also exploits a method called bagging, which eventually lets each decision tree randomly extract the sample dataset without any predefined rules in order to have different variants of trees as an output, and achieve better accuracy while maintaining minimum variance. Moreover, the algorithm utilizes feature randomness that further helps to increase diversification. The RF is calculated as given in equation 1.

$$\boldsymbol{RFfi_i} = \frac{\left(\boldsymbol{\Sigma_{\{j \in all\ trees\}} normfi_{\{ij\}}}\right)}{\boldsymbol{T}} \quad (1)$$

Here, RFfi_i = The importance of feature i is determined by aggregating its contribution across all trees. normfi_{ij} = the feature i is further normalized in a tree called j,{T} = represents the count of trees

3.4.2. Logistic Regression (LR)

Logistic Regression is considered another prominent method in machine learning, widely recognized as a supervised algorithm used for classification tasks [25]. This technique is based on linear regression, with its primary feature being that LR uses an activation function to limit outputs to a range between 0 and 1. The value obtained signifies the probability of a particular output class, which also lies between 0 and 1. The final computed probability is then used to create a set of discrete values. Furthermore, logistic regression is applicable not only for binary classification but also for multiclass classification scenarios. The mathematical formulation of LR is illustrated in equation 2.

$$\boldsymbol{P} = \frac{\mathbf{1}}{\{\mathbf{1} + \boldsymbol{e}^{\{-(\boldsymbol{a}+\boldsymbol{bX})\}}\}} \quad (2)$$

As per equation 2, {P} represents the probability of having 1, while *e* shows the base of logarithms, {a} and {b} are known as parameters, and {X} is called an independent variable.

3.4.3. Support Vector Machine (SVM)

Support Vector Machine falls under the category of supervised machine learning algorithms that are eventually used for the tasks of classification as well as regression. The algorithm works on the phenomenon of mapping data points in a space having n-dimensions. As a result, the classification task occurs by separating the plotted data points with the help of a hyperplane [26]. Similarly, SVM computes the distance between the data points having opposing labels, and these are known as support vectors. With the help of these calculated data points, a margin is constructed that improves the model's generalization. These margins are divided into two categories such as hard margin and soft margin. The margins cause overfitting along with misclassification when generalizing the model. The equation for the support vector machine is represented in Equation 3.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p \geq 0 \quad (3)$$

3.4.4. K-nearest neighbors (KNN)

K-nearest neighbor is another commonly used supervised machine learning method that is used for classification and regression. The algorithm predicts the class by computing the distance between the data points and the training point [27]. The computed distance between data points in a dimensional space is usually exploited with the help of Euclidean distance or Minkowski distance, as shown in the equation 4.

Once the distance is calculated, the nearest K points are chosen, and the prediction is generated by assessing the classes that occur the most frequently among these K points. The most appropriate value gets chosen depending on the size of the dataset.

$$\boldsymbol{d} = \boldsymbol{sqrt}\left(\boldsymbol{sum}_{\{\boldsymbol{i}=\mathbf{1}\}}^{\{\boldsymbol{k}\}(\boldsymbol{x_i}-\boldsymbol{y_i})^2}\right) \quad (4)$$

3.4.5. Decision Tree

A decision tree is considered one of the effective machine-learning approaches that has been utilized for both predictions. Its format is akin to that of a tree, with each branch signifying a feature from the dataset. These branches accordingly assist in ascertaining the outcome of a data point that is being evaluated. During the configuration, each leaf node signifies a class label that needs to be predicted. To build a decision tree, the entropy as well as information gained for every feature are computed in comparison to the target variable to ascertain the significance of feature nodes within the tree. Finally, once the tree is constructed, predictions for new data points are computed based on tracing the path from the

root nodes up to the leaf nodes. The mathematical expressions for computing entropy are presented in Equation 5.

$$\boldsymbol{E(S) = -\Sigma_{c} - pilog2pi} \quad (5)$$

3.4.6. Naive Bayes

Naive Bayes is a simple but powerful probabilistic model used for classification purposes. It is based on Bayes' Theorem, which calculates the probability of a hypothesis given the observed data. The approach assumes that the features are independent of each other, a condition that is often not met in real-world datasets; nevertheless, this 'naive' assumption simplifies calculations and yields surprisingly effective results in practical applications [28]. Naive Bayes is especially useful for text classification, such as spam filtering or sentiment analysis, where it swiftly assesses the probabilities of categories based on the occurrence of words. Despite its simplicity, it performs well with large datasets and is robust in handling noise. The calculation of Naive Bayes is represented in Equation 6.

$$P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C) \quad (6)$$

3.5. Deep Learning Models

The following section elaborates on the details of three deep learning algorithms that have been applied in the proposed study on the Spotify review dataset to evaluate the effectiveness of each of the aforementioned algorithms.

3.5.1. Recurrent Neural Network

The use of repeated neural networks (RNN) in data sets such as Spotify Reviews shows the ability to capture a consistent dependency of the text data. RNN is especially effective in handling consistent information because it supports internal memory for storing previous input. This helps in the context of modeling text data and long-term dependencies. In this study, the SPOTIFY review data set containing user reviews and labels has been used to ensure homogeneous input sequences using tokenization and gas kit methods. The proposed model uses a single-layer RNN with 64 repeated units, along with layers designed for the falling layer for expression and regulation of word vectors. The learning process was able to converge faster using binary losses from cross-entropy and the Adam Optimizer. We assessed the effects of the model using the evaluation indicators, including accuracy, review, and F_1 accuracy, providing comprehensive information on the effectiveness of classification. According to the experimental results, RNN is suitable for atmosphere classification because it effectively captures the template in a user review and shows competitive accuracy and balanced performance between the metrics [29].

3.5.2. *Long Short-Term Memory*

Long short-term memory (LSTM) is an advanced type of recurrent neural network (RNN) aimed at addressing the challenges of gradient vanishing and capturing long-term dependencies in sequential data. In this research, both single-layer and multi-layer LSTM structures were utilized on Spotify review data to assess their impact on mood classification. The single-layer LSTM model provided a basic line for performance using the ability to maintain the situation information into another sequence. In order to improve the extraction of the signs, a deeper three-story LSTM model was implemented to provide training for hierarchical expressions. Each tier of multilayer architecture is processed by sequential data in different levels of abstraction, improving the ability of the model to detect complex templates. To prevent overfitting, a drop-down layer was added between the LSTM layer to control. Both models were evaluated using accurate response, F_1 estimated value, and accuracy indicator, and the results show that the multi-layer LSTM surpassed a single-layer LSTM. Especially when subtle templates are captured with text data. This conclusion emphasizes the effect of a deeper LSTM architecture on the task of analyzing the text atmosphere [30]. In the SPOTIFY review, the LSTM single-layer network effectively models the consistent template of the text data. This model uses a built layer to create a dense vector representation and sequences the LSTM layer with 64 units to capture dependence. In the case of Regularization, the falling layer is added to reduce the risk of experience. The double losses of horizontal entropy and Adam Optimizer are used to teach models. Evaluation indicators such as accuracy, review, F_1 evaluation, and accuracy show the ability to effectively analyze moods in user reviews by emphasizing the performance of the model [31].

3.5.3. Gated Recurrent Neural Network

The protective recurrence unit (GRUS) is a deformation of the repeated nerve network (RNN), which simplifies the architecture of the long-term memory network (LSTM) and maintains the ability to simulate

the long-term dependency in continuous data. GRUS uses a string mechanism to adjust the flow of information, so the calculation is more effective than LSTM. In this study, a single and multilayer GRU model has been implemented to analyze the atmosphere of Spotify Reviews. The single-layer GRU was used as a default level, using the ability to effectively handle consistent dependencies. To capture more complex patterns, three layers of GRU architecture can be developed to hierarchical objects in the layer. The complex layer is included between the GRU layers, and improves the generalization [32]. Evaluation indicators, including accuracy, recall, precision, and F_1 score, showed that multi-layer GRUs surpassed a single-layer model, especially when seizing complex dependencies of text data. This result emphasizes the appropriateness of GRU, especially deeper architectures, for tasks that classify texts that calculate efficiency, and model sequences [33].

3.6. Transformer-based Models

The model based on the transformer is the latest artificial intelligence system designed to process and create human text based on a huge amount of data. They use a transformer architecture that analyzes the relationship between words in sentences using the same mechanisms to provide contextual predictions and answers [34]. Models such as GPT and BERT [35] are commonly deployed for applications, such as in the domain of research, business, and communication, by using these architectures to respond to text, translation, summary, and questions.

3.6.1. BERT-Based Model

BERT is an innovative model that is widely used in the domain of NLP and is presented by Google, and is a rotation of a method for the machine to understand and handle text. Unlike the existing model, BERT can use the bucketed approach to the context of the words of the previous text and subsequent text at the same time [36]. Understanding this deep situation is very effective for a wide range of NLP work, including mood analysis, answers to questions, and language translations. The BERT model is prepared in a wide range of buildings, such as Wikipedia, and sets a thin-tailored new test for certain tasks to show a variety of tests and strengths in complex verbal nuances [37].

3.6.2. BERT Large Model

With a deeper design that enhances its ability to understand complex linguistic structures, BERT Large is a more advanced and potent version of the BERT model. The model's 340 million parameters—which is significantly higher than the 110 million parameters in BERT Base—come from 24 transformer layers, a hidden layer of 1024, and around 16 attention heads. By identifying intricate linguistic features, this improved capability helps BERT Large perform exceptionally well on a variety of natural language processing tasks, such as reading comprehension and question answering. However, as these benefits come with higher computational costs, longer training times, and a greater chance of overfitting, it works best with big datasets and scenarios with plenty of computer power [38].

3.7. Evaluation Metrics

A range of assessment metrics is utilized by researchers to evaluate the efficacy of algorithms in various applications. These evaluation metrics allow researchers to analyze and compare the overall performance of the algorithms, thereby facilitating the decision-making process. The proposed method is currently being evaluated using the following metrics:

3.7.1. Accuracy

One of the most important and crucial evaluation metrics that is mostly considered for measuring the effectiveness of the model is Accuracy. Accuracy plays an important role in the classification of algorithms, mainly for the domains of Machine Learning and Deep Learning [39-46]. The ratio of accurate predictions to the total number of input data samples is its primary definition. Here, genuine positives are denoted by TP, true negatives by TN, false positives by FP, and false negatives by FN.

3.7.2. Recall

An alternative metric for evaluating the performance of classification algorithms is known as Recall [40]. By counting the number of accurately detected positives in the test dataset, this statistic evaluates the true positive rate.

3.7.3. Precision

Precision serves as another metric for evaluating the performance of classification algorithms. It assesses the proportion of accurately identified positives within our test dataset [41]. The formula to determine precision is the ratio of true positive cases to the total number of positives that have been

identified in the sample dataset. A classification algorithm will achieve higher precision if it accurately predicts positive samples effectively.

3.7.4. F_1 Score

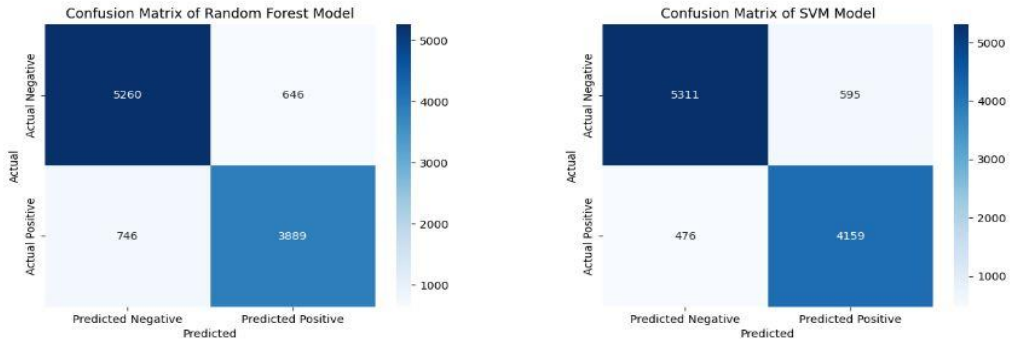
The F_1 score serves as a metric to evaluate the performance for classification tasks in the field of Machine Learning. It is specifically formulated to address issues related to class imbalance [42]. The calculation of the F_1 Score is based on precision and recall. The method for determining the F_1 Score is given by computing the harmonic mean of these two values. When both Precision and Recall are elevated, the F_1 Score will also increase. On the other hand, if both are low, the resulting score will also be low. If one metric is high while the other is low, the result will be an intermediate value.

4. Result and Discussion

The section elaborates on the results obtained after exploiting different machine learning, deep learning, and large language models using the Spotify Review dataset, along with a detailed description of the algorithms.

4.1. Machine Learning Results using TF-IDF, PoS Tagging, and BoW Features

This subsection presents the outcome of the machine-learning models along with their analysis with respect to different feature extraction techniques in order to analyze the results against the positive as well as negative classes. As can be seen earlier, Table 2 provides the performance matrices for various machine learning models while applying different feature engineering techniques. Among the seven machine learning classifiers, the results show that after applying TF-IDF, RF, and SVM, a good performance was achieved, in terms of accuracy, when compared to all other classifiers, as illustrated in Figure 4. Similarly, as shown Figure 5, regarding POS tagging and BoW, SVM outperforms the other state-of-the-art classifiers by achieving 90% accuracy.



(a) Random Forest (b) Support Vector Machine

Figure 4. Confusion matrix of the top-performing Machine Learning classifier using TF-IDF

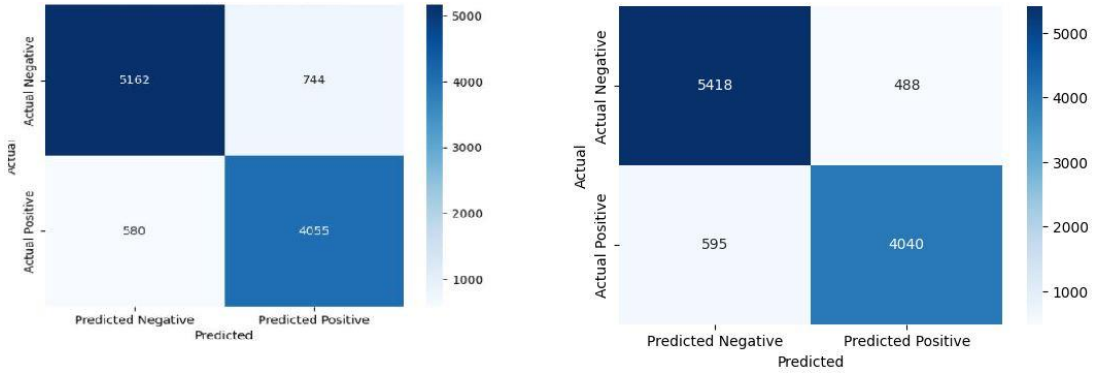


Figure 5. Confusion Matrix of SVM Classifier using Bag of Words (L.H.S) and PoS Tagging (R.H.S)

Table 2. Performance Metrics per class for Various Machine Learning Models with Different Feature Engineering Techniques

Model	Feature Extraction	Accuracy	Precision	Recall	F_1-Score

SVM	BoW	0.87	0.85	0.80	0.86
	TF-IDF	0.87	0.84	0.88	0.82
	POS	0.90	0.86	0.89	0.87
RF	BoW	0.86	0.85	0.72	0.78
	TF-IDF	0.87	0.80	0.61	0.70
	POS	0.86	0.84	0.77	0.83
NB	BoW	0.86	0.82	0.85	0.81
	TF-IDF	0.86	0.79	0.69	0.73
	POS	0.86	0.85	0.78	0.84
LR	BoW	0.86	0.88	0.85	0.86
	TF-IDF	0.87	0.88	0.79	0.82
	POS	0.87	0.87	0.82	0.83
DT	BoW	0.79	0.75	0.83	0.74
	TF-IDF	0.79	0.78	0.74	0.80
	POS	0.79	0.73	0.81	0.72
KNN	BoW	0.65	0.63	0.74	0.42
	TF-IDF	0.66	0.58	0.62	0.36
	POS	0.46	0.86	0.99	0.62

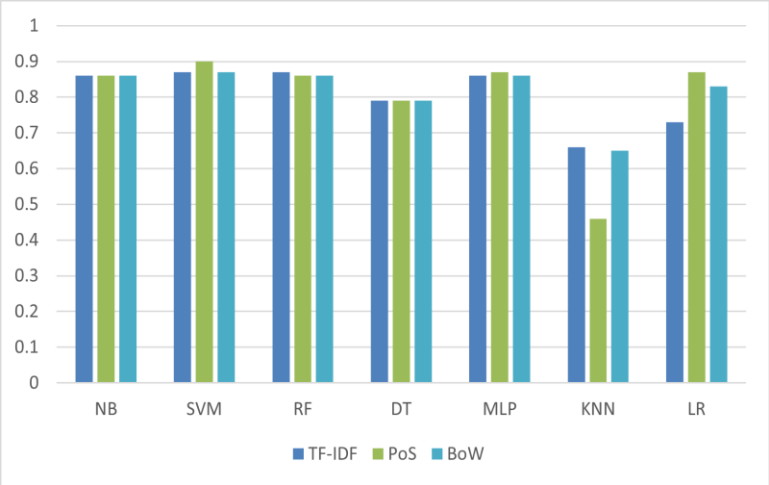


Figure 6. Comparison of Machine Learning models using TF-IDF, Bag of Words, and PoS tagging

4.2. Deep Learning Algorithm Results

After exploiting different variants of three prominent deep learning algorithms, namely Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Neural Network (GRU), using 1 layer and 3 layers, respectively. The detailed results of deep learning algorithms on the Spotify review dataset are demonstrated in Table 3. As can be seen, among all the algorithms, LSTM with 1 layer outperforms all other state-of-the-art deep learning technologies in terms of accuracy, precision, recall, and F_1-score. The results in terms of accuracy have been visualized in Figure 7.

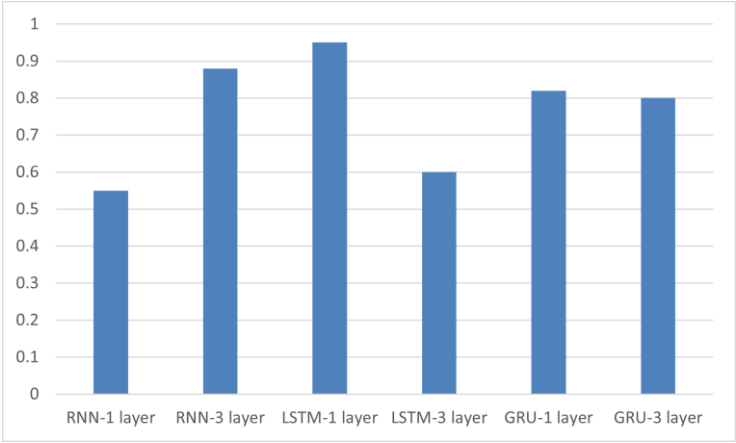


Figure 7. Comparison between Deep learning models in terms of Accuracy

To further understand the contribution of each feature extraction method, an ablation-style comparison was conducted conceptually by analyzing model performance across Bag of Words (BoW), TF-IDF, and Part-of-Speech (POS) tagging. The results indicate that models trained using BoW and TF-IDF representations achieved relatively higher accuracy compared to those relying solely on POS-based features. This suggests that frequency- and distribution-based representations capture sentiment-bearing terms more effectively, while POS features primarily contribute to syntactic context. Although a formal ablation study isolating these effects was not performed, the observed trends highlight the complementary nature of these methods.

4.3. Transformer-based Model Results

This result section provides the results obtained after implementing the Bert-uncased Model and Bert-large model on the Spotify review dataset and indicates that these models tend to perform better than most of the deep learning models, achieving 92% and 94%, respectively.

Table 3. Results of Deep Learning and Transformer Model

Model	Accuracy	Precision	Recall	F1-Measure
RNN-1	0.55	0.99	0.85	0.17
RNN-3	0.88	0.84	0.91	0.87
LSTM-1	0.95	0.95	0.92	0.93
LSTM-3	0.60	0.73	0.67	0.77
GRU-1	0.82	0.78	0.81	0.79
GRU-3	0.80	0.78	0.76	0.77
BERT-based Model	0.92	0.87	0.85	0.87
BERT-large Model	0.94	0.90	0.86	0.89

5. Conclusion and Future Work

Due to drastic advancements in automation and a rapid surge in the growth of music streaming services, there is a need to develop such models and evaluate their performance to improve the user experience and satisfaction for music streaming platforms such as Spotify. Spotify, among other competitors, has gained significant popularity among music listeners over the past few years, and this fame makes it stand out as one of the most popular choices among users, because of its wide variety of music library with options such as personalized playlists, as well as a user-friendly interface. Though all other platforms have their own strengths and shortcomings, Spotify has proven to be the top runner among all. The proposed approach presents a wide-ranging analysis of Spotify user reviews while leveraging the conventional state-of-the-art ML, DL, and transformer-based models in order to classify the user sentiments effectively and accurately.

The conventional ML algorithms, such as SVM, RF, NB, LR, DT, and KNN, have been deployed to assess the performance of each of the models and then make a comparative analysis among all. The results demonstrated that SVM outperforms all other state-of-the-art models by achieving an accuracy of 90%. Similarly, on the other hand, several DL models and their variants were deployed, such as RNN, CNN, and LSTM, to evaluate the efficacy of each of them for the task of sentiment classification. The results

showed that LSTM with 1 layer achieved the highest accuracy of 95% while RNN with 1 layer obtained a 99% precision rate. Alongside the two widely known transformer-based models, such as BERT and its variant BERT-large, were employed, among which BERT-large exhibited superior performance by obtaining an accuracy of 94%.

The proposed study revealed that users experienced high satisfaction related to the quality of the music and the playlist recommendation, while apprehensions were raised, such as related to the cost of subscription, app performance issues, and ad interruptions, mainly in the free version. Concludingly, the DL models efficiently apprehended the contextual meanings within reviews, while Large Language Models (LLMs) showed notable performance in mining insights from user reviews that are usually complex data patterns. Even though the LSTM and BERT-large models achieved superior performance, they demanded significantly higher computational resources and training time, equating to traditional algorithms such as SVM.

Notwithstanding the promising outcomes, this investigation is subject to several limitations. Firstly, it lacks a comprehensive ablation analysis to disentangle the impact of each feature extraction technique. Secondly, the scope of the dataset is confined to English-language Spotify reviews, which may impose constraints on the generalizability of the findings. Moreover, the aspects of computational expense and model interpretability were not scrutinized in detail, which are essential factors for practical implementation. Furthermore, considering the dataset's slight imbalance (56% negative, 44% positive), there exists a possibility that an inclination toward negative sentiment could have skewed the results.

In order to provide more detailed insights, we anticipate improving the analysis in future work by integrating multilingual evaluations and investigating feelings unique to a certain region. The size of the dataset, which included 51,473 reviews, is one of the study's limitations. Smaller subsets can be produced by splitting these evaluations into several subjects, which might affect the multiplicity of the topics. Subsequently, many users describe numerous features in a single review; we also aim to handle overlapping subjects in user reviews. Future research will emphasize inspecting sentiment patterns over time and integrating data from other sources, such as comments on social media.

Funding: This research received no funding.
Data Availability Statement: Data is publicly available on Kaggle here.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yu, S. (2024). Case Study: Artificial Intelligence in Spotify. *SSIE Winter Research*.
2. M. Giacomo, "Spotify: A Strategic Analysis of a Digital Music Streaming Industry Leader," 2024.
3. P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, 2020, pp. 99–111.
4. O. Oyebode, F. Alqahtani, and R. Orji, "Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews," IEEE Access, vol. 8, pp. 111141–111158, 2020.
5. S. Stradowski and L. Madeyski, "Industrial applications of software defect prediction using machine learning: A business-driven systematic literature review," Inf Softw Technol, vol. 159, p. 107192, 2023.
6. M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine Learning for industrial applications: A comprehensive literature review," Expert Syst Appl, vol. 175, p. 114820, 2021.
7. G. Apruzzese et al., "The role of machine learning in cybersecurity," Digital Threats: Research and Practice, vol. 4, no. 1, pp. 1–38, 2023.
8. S. Chen, Y. Zhang, and Q. Yang, "Multi-task learning in natural language processing: An overview," ACM Comput Surv, vol. 56, no. 12, pp. 1–32, 2024.
9. M. M. Raharjo and F. Arifin, "Machine learning system implementation of education podcast recommendations on spotify applications using content-based filtering and tf-idf," Elinvo (Electronics, Informatics, and Vocational Education), vol. 8, no. 2, pp. 221–230, 2023.
10. T. K. Bushra, K. Saha, A. H. Mulki, S. S. Khan, and A. Binta Amzad, "Recognizing sentimental emotions in text by using Machine Learning," Brac University, 2022.
11. G. Björklund, M. Bohlin, E. Olander, J. Jansson, C. E. Walter, and M. Au-Yong-Oliveira, "An Exploratory Study on the Spotify Recommender System," in World Conference on Information Systems and Technologies, 2022, pp. 366–378.
12. D. Hendri, D. Nadha, F. K. Basri, M. F. Wajdi, and N. Nadhirah, "Comparation of Decision Tree Algorithm, Naive Bayes, K-Nearest Neighbords on Spotify Music Genre," IJATIS: Indonesian Journal of Applied Technology and Innovation Science, vol. 1, no. 1, pp. 47–53, 2024.
13. F. Khan et al., "Effect of feature selection on the accuracy of music popularity classification using machine learning algorithms," Electronics (Basel), vol. 11, no. 21, p. 3518, 2022.
14. N. Pelchat and C. M. Gelowitz, "Neural network music genre classification," Canadian Journal of Electrical and Computer Engineering, vol. 43, no. 3, pp. 170–173, 2020.
15. M. B. Saputro and A. Alamsyah, "Comparison of Naive Bayes Classifier and K-Nearest Neighbor Algorithms with Information Gain and Adaptive Boosting for Sentiment Analysis of Spotify App Reviews," Recursive Journal of Informatics, vol. 2, no. 1, pp. 37–44, 2024.
16. A. and others Triyono, Agung and Faqih, "Implementation of the Naive Bayes Method in Sentiment Analysis of Spotify Application Reviews," Journal of Artificial Intelligence and Engineering Applications (JAIEA), 2025.
17. A. Triyono, A. Faqih, and G. Dwilestari, "Journal of Artificial Intelligence and Engineering Applications Implementation of the Naive Bayes Method in Sentiment Analysis of Spotify Application Reviews," vol. 4, no. 2, pp. 2808–4519, 2025, [Online]. Available: https://ioinformatic.org/
18. G. Eser and C. Sahin, "Sentiment Analysis and Rating Prediction for App Reviews Using Transformer-based Models," International Journal of Advanced Natural Sciences and Engineering Researches, vol. 4, no. 4, pp. 372–379, 2024, doi: 10.5281/zenodo.12731064.
19. J. Ram\'\irez and M. J. Flores, "Machine learning for music genre: multifaceted review and experimentation with audioset," J Intell Inf Syst, vol. 55, no. 3, pp. 469–499, 2020.
20. Nur Istiqamah, "Spotify Podcast Integration in Listening Learning: Impact on Students' Digital Literacy," Jurnal QOSIM Jurnal Pendidikan Sosial & Humaniora, vol. 3, no. 2, pp. 698–705, 2025, doi: 10.61104/jq.v3i2.1105.
21. C. N. N. Architectures et al., "The Indonesian Journal of Computer Science," vol. 14, no. 2, pp. 2121–2139, 2025.
22. Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based BERT model in sentiment analysis," Applied and Computational Engineering, vol. 71, no. 1, pp. 14–20, 2024, doi: 10.54254/2755-2721/71/2024ma.
23. P. Bakhsh, M. Ismail, M. A. Khan, M. Ali, and R. A. Memon, "Optimisation of Sentiment Analysis for E-Commerce," VFAST Transactions on Software Engineering, vol. 12, no. 3, pp. 243–262, 2024, doi: 10.21015/vtse.v12i3.1907.

24. J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," Brief Bioinform, vol. 24, no. 2, p. bbad002, 2023.
25. G. Aliman et al., "Sentiment analysis using logistic regression," Journal of Computational Innovations and Engineering Applications, vol. 7, no. 1, pp. 35–40, 2022.
26. A. Roy and S. Chakraborty, "Support vector machine in structural reliability analysis: A review," Reliab Eng Syst Saf, vol. 233, p. 109126, 2023.
27. M. Suyal and P. Goyal, "A review on analysis of k-nearest neighbor classification machine learning algorithms based on supervised learning," International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 43–48, 2022.
28. E. M. K. Reddy, A. Gurrala, V. B. Hasitha, and K. V. R. Kumar, "Introduction to Naive Bayes and a review on its subtypes with applications," Bayesian reasoning and gaussian processes for machine learning applications, pp. 1–14, 2022.
29. A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D, vol. 404, p. 132306, 2020.
30. R. C. Staudemeyer and E. R. Morris, "Understanding LSTM–a tutorial into long short-term memory recurrent neural networks," arXiv preprint arXiv:1909.09586, 2019.
31. Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," Neural Comput, vol. 31, no. 7, pp. 1235–1270, 2019.
32. F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," arXiv preprint arXiv:2305.17473, 2023.
33. T. H. Lee et al., "Comparative Analysis of 1D–CNN, GRU, and LSTM for Classifying Step Duration in Elderly and Adolescents Using Computer Vision," International Journal of Robotics and Control Systems, vol. 5, no. 1, pp. 426–439, 2025.
34. A. Vaswani, "Attention is all you need," Adv Neural Inf Process Syst, 2017.
35. P. Ganesh et al., "Compressing large-scale transformer-based models: A case study on bert," Trans Assoc Comput Linguist, vol. 9, pp. 1061–1080, 2021.
36. F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," Artif Intell Rev, vol. 54, no. 8, pp. 5789–5829, 2021.
37. M. V Koroteev, "BERT: a review of applications in natural language processing and understanding," arXiv preprint arXiv:2103.11943, 2021.
38. R. Anil, B. Ghazi, V. Gupta, R. Kumar, and P. Manurangsi, "Large-scale differentially private BERT," arXiv preprint arXiv:2108.01624, 2021.
39. A. Tharwat, "Classification assessment methods," Applied computing and informatics, vol. 17, no. 1, pp. 168–192, 2021.
40. J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," Evol Intell, vol. 15, no. 3, pp. 1545–1569, 2022.
41. Ž. Vujović and others, "Classification model evaluation metrics," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 599–606, 2021.
42. D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," Mach Learn, vol. 110, no. 3, pp. 451–456, 2021.
43. Z. Awais *et al.*, "ISCC: Intelligent Semantic Caching and Control for NDN-Enabled Industrial IoT Networks," in *IEEE Access*, vol. 13, pp. 169881-169898, 2025, doi: 10.1109/ACCESS.2025.3614984.
44. Zubair, M.; Hussain, M.; Albashrawi, M.A.; Bendechache, M.; Owais, M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. Computer Methods and Programs in Biomedicine. **2025**, *272*, 109014. https://doi.org/10.1016/j.cmpb.2025.109014.
45. Hussain, M., Chen, C., Hussain, M. *et al.* Optimised knowledge distillation for efficient social media emotion recognition using DistilBERT and ALBERT. *Sci Rep* **15**, 30104 (2025). https://doi.org/10.1038/s41598-025-16001-9.
46. Zubair, M., Owais, M., Hassan, T. *et al.* An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. *Sci Rep* **15**, 13087 (2025). https://doi.org/10.1038/s41598-025-97256-0.