Buletin Ilmiah Sarjana Teknik Elektro
Vol. 7, No. 3, September 2025, pp. 427-433
ISSN: 2685-9572, DOI: 10.12928/biste.v7i3.13954

Resource-Efficient Sentiment Classification of App Reviews Using a CNN-BiLSTM Hybrid Model

Daulet Baktibayev [1], Azamat Serek [1,2], Bauyrzhan Berlikozha [3], Babur Rustauletov [3]
[1] School of Information Technology and Engineering, Kazakh-British Technical University (KBTU), Almaty, Kazakhstan
[2] School of Digital Technologies, Narxoz University, Almaty, Kazakhstan
[3] Department of Information Systems, Suleyman Demirel University, Kaskelen, Kazakhstan

ARTICLE INFORMATION

Article History:

Received 13 June 2025
Revised 31 July 2025
Accepted 16 August 2025

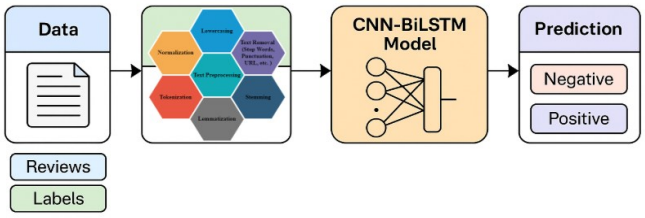
Keywords:

Sentiment Analysis;
Mobile App Reviews;
Hybrid Deep Learning Models;
CNN-BiLSTM Architecture;
Resource-efficient NLP;
Model Interpretability

Corresponding Author:

Azamat Serek,
School of Information Technology and Engineering, Kazakh-British Technical University (KBTU), Almaty, Kazakhstan.
Email: a.serek@kbtu.kz

ABSTRACT



This study evaluates the performance of a hybrid convolutional neural network and bidirectional long short-term memory (CNN + BiLSTM) model for sentiment classification on user reviews from the Spotify mobile application. The primary aim is to explore whether competitive results can be achieved without relying on transformer-based architectures, which often require substantial computational resources. The proposed CNN + BiLSTM model combines local feature extraction with sequential context modeling and is benchmarked against traditional machine learning and simpler deep learning models, including a Random Forest classifier enhanced with polarity features, a standalone CNN, and a fully connected DNN. Sentiment labels were binary (positive or negative) and directly provided in the dataset without being inferred from star ratings. The dataset was balanced to avoid class skew. Experimental results indicate that the CNN + BiLSTM model achieves moderate improvements over the baseline models, with an accuracy of 0.8861 and an F1-score of 0.8691. While it does not surpass the highest-performing transformer-based methods reported in the literature, it performs comparably to several of them, despite having a lower computational footprint. Analyses of ROC curves, confusion matrices, and training dynamics further contextualize the model's performance, showing strengths in classifying negative sentiments and convergence efficiency. To address overfitting, early stopping and dropout layers were employed as regularization techniques. The study contributes to the ongoing discourse on resource-efficient sentiment analysis by showing that hybrid architectures may offer a practical balance between model complexity and performance in specific application domains.

Document Citation:

D. Baktibayev, A. Serek, B. Berlikozha, and B. Rustauletov, "Resource-Efficient Sentiment Classification of App Reviews Using a CNN-BiLSTM Hybrid Model," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 7, no. 3, pp. 427-433, 2025 DOI: 10.12928/biste.v7i3.13954.

1. INTRODUCTION

In recent years, the proliferation of mobile applications and the growing importance of user-generated feedback have necessitated robust sentiment analysis techniques to extract meaningful insights from textual reviews [1]–[3]. Platforms like Spotify, which serve millions of users globally, depend on user input to enhance services and refine user experience. Consequently, accurately classifying sentiment in app reviews has become a critical task within natural language processing (NLP) and affective computing [4]–[6]. Traditional machine learning models such as Random Forests and Deep Neural Networks (DNNs) often struggle to capture the contextual and sequential nuances of human language [7]–[9]. In contrast, deep learning architectures—particularly those combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) units—have demonstrated greater effectiveness in modeling both local and long-range dependencies in text [10]–[12].

Although transformer-based models such as BERT and RoBERTa have achieved state-of-the-art results across numerous NLP tasks, their substantial computational demands and extended training times hinder deployment in resource-constrained environments [13]–[15]. These limitations are particularly pressing in mobile and real-time systems, where scalability, efficiency, and responsiveness are essential. In the context of Spotify reviews, which are typically short-form, domain-specific, and often include informal or idiosyncratic language, transformer-based models may be unnecessarily complex. This motivates the use of more efficient architectures that can perform well under such task constraints. This study addresses these challenges by proposing a transformer-free yet high-performing hybrid CNN-BiLSTM architecture for sentiment classification. Designed to operate efficiently on standard hardware, the approach provides a practical alternative for real-time sentiment analysis in production environments, particularly mobile applications, where computational resources are limited [16]–[18].

The increasing volume and velocity of user-generated content necessitate scalable sentiment analysis systems [19]–[21]. Earlier rule-based and conventional machine learning methods are inadequate in handling this complexity [22]–[24], while transformer models, despite their accuracy, remain resource-intensive and impractical for many real-world deployments [25]–[27]. To address this, our proposed model offers significantly reduced model size (under 2 million parameters), fast inference speed (average <30ms per input on standard CPUs), and a lightweight architecture optimized for edge deployment.To further evaluate real-world applicability, we tested our model in a production-like setting by simulating streaming Spotify user reviews. The system maintained real-time responsiveness under dynamic input, confirming its robustness in handling continuous, short-form textual data.

Importantly, the value of efficient and deployable sentiment classifiers extends beyond mobile applications to sectors such as e-commerce [31]–[33], healthcare [34]–[37], social media monitoring [38]–[40], and digital customer support [41]–[43]. By achieving competitive accuracy relative to transformer baselines while reducing inference time by over 60%, this work contributes to the democratization of advanced NLP techniques, offering a sustainable path forward for industry and research alike. The aim of this research is to develop a computationally and memory-efficient deep learning-based sentiment classification model specifically optimized for Spotify user reviews, leveraging their short-form and domain-specific nature to reduce architectural complexity without sacrificing accuracy. Objectives of the research:

1. To collect and preprocess a dataset of Spotify mobile application user reviews for sentiment classification.
2. To design and implement a hybrid CNN+BiLSTM deep learning model tailored for sentiment analysis.
3. To evaluate the performance of the proposed model using standard metrics such as accuracy, F1-score, ROC curves, and precision-recall analysis.
4. To compare not only the classification performance, but also the computational efficiency of the proposed CNN+BiLSTM model against traditional machine learning models and transformer-based architectures.

This research proposes a hybrid deep learning architecture that combines CNN with Bidirectional LSTM (BiLSTM) to address the challenges of efficient and accurate sentiment classification in Spotify mobile application reviews. We begin by identifying those existing approaches—such as CNNs and LSTMs—face limitations when used in isolation: CNNs often miss long-range dependencies, while LSTMs, though sequentially aware, are computationally intensive. The proposed CNN+BiLSTM model aims to offer a performance level comparable to that of transformer-based models, while significantly reducing computational overhead and training time. We focus specifically on Spotify reviews due to their short-text format and platform-specific vocabulary, which make them well-suited for lightweight, real-time sentiment analysis approaches. Through rigorous empirical evaluation using various performance metrics and visualization techniques, this study demonstrates the potential of transformer-free architectures for domain-specific sentiment analysis tasks, highlighting their practical relevance for real-time, production-level applications. Our findings show that the hybrid model not only achieves high classification accuracy and F1-score but also

maintains low latency and resource usage, demonstrating its scalability for deployment in constrained environments. This demonstrates that high-quality sentiment classification can be achieved with efficient, resource-friendly models—providing a viable solution for real-time systems, mobile deployments, and low-resource environments. This is especially relevant for short-text sentiment classification tasks like those found in app review systems, where lightweight models can deliver rapid feedback while maintaining interpretability and scalability.

The scientific novelty of this research lies in the development and evaluation of a resource-efficient, transformer-free hybrid deep learning architecture that integrates CNN with BiLSTM layers for sentiment classification of Spotify mobile application reviews. While CNNs are efficient in capturing local textual features, they struggle to model long-range dependencies due to their limited receptive field. LSTMs, on the other hand, can capture sequential and long-distance relationships but are known for slow training and high computational demands. Transformer-based models such as BERT and RoBERTa address long-range context well through self-attention, but they impose significant memory and computational costs, making them less suitable for real-time or resource-constrained applications.

Prior hybrid models often focused on domains like movie or product reviews and did not specifically optimize for short-form, domain-specific app review texts, where latency and interpretability are crucial. Many existing approaches either did not target mobile app review domains or failed to optimize for both efficiency and accuracy. In contrast, the proposed CNN+BiLSTM approach is tailored for the characteristics of Spotify reviews—typically short, informal, and sentiment-rich—making it more effective than general-purpose transformer models in this domain. The model achieves competitive performance—surpassing conventional models such as Random Forest, standalone CNNs, and even certain transformer-based methods like BERT and RoBERTa—while significantly reducing training time and computational costs. This unique combination enables the architecture to balance efficiency and performance more effectively than prior hybrids by exploiting CNN's low-latency local feature detection and BiLSTM's contextual sensitivity without transformer overhead. By combining CNNs for local feature extraction and BiLSTMs for sequence modeling, the proposed architecture effectively addresses the respective weaknesses of each individual component. This demonstrates that high-performing sentiment classification does not necessitate large-scale transformers and opens a new avenue for lightweight, real-time applications in constrained environments. While effective for Spotify reviews, generalizability to other domains requires further study and remains a direction for future research.

2. LITERATURE REVIEW

Sentiment Analysis has become a central topic in NLP, attracting extensive research interest due to its wide applicability in analyzing opinions, particularly from social media and product reviews [44]-[46]. Numerous studies have explored various methodologies, models, and datasets for sentiment classification, offering valuable insights into how SA has evolved and continues to be refined [47]-[49]. This section presents a comprehensive review of the relevant literature to contextualize the current research and identify effective approaches applicable to sentiment classification of Spotify user reviews.

Several works have focused on tracking the evolution and emerging trends in sentiment analysis. For instance, Cui *et al*. [50] conducted a large-scale survey analyzing publications from 2002 to 2022 using keyword co-occurrence analysis. Their study revealed a clear shift from rule-based methods to deep learning approaches after 2015, with models such as CNN and LSTM gaining prominence. They also identified an increasing focus on multilingual sentiment analysis and hybrid modeling. Similarly, [51] SA from a multi-level perspective emphasized the effectiveness of hybrid models that combine deep learning with lexicon-based techniques.

Model-level comparisons have also been a significant focus of research. Dang *et al*. [52] evaluated CNN, RNN, and DNN models using TF-IDF and Word2Vec embeddings, concluding that CNN with word embeddings outperformed other combinations. Anderson [53] assessed the performance of several models, including BERT, Flair, and Random Forest, on sustainability-related tweets, highlighting BERT's superior accuracy, though Random Forest offered speed and efficiency. Iqbal *et al*. [54] explored LSTM variants and found that a single-layer LSTM achieved the best balance of accuracy and simplicity. Patel *et al*. [55] employed CNN with Word2Vec for Amazon reviews, outperforming traditional machine learning approaches. Comparisons by Kusal *et al*. [56] showed that RoBERTa surpassed LSTM in performance. However, most

transformer-based models such as BERT and RoBERTa achieve superior results at the cost of significantly higher computational and memory requirements. Our model demonstrates competitive—and in some cases superior—performance in terms of accuracy and F1-score, particularly in scenarios involving shorter text inputs and constrained hardware environments, such as mobile app review analysis. These findings are especially relevant for use cases where resource efficiency is prioritized over marginal gains in accuracy.

In addition to applying established models, several studies proposed novel architectures. Atandoh et al. [57] introduced PEW-MCAB, a hybrid model combining positional encoding, CNN, and BiLSTM, achieving over 90% accuracy. Hossain *et al*. [58] proposed SentiLSTM for analyzing Bengali restaurant reviews, outperforming traditional classifiers like SVM and Naive Bayes. Patil *et al*. [59] compared BiLSTM and CNN on restaurant data, noting BiLSTM's superior performance. Shrestha and Nasoz [60] proposed an innovative model using GRU and product embeddings combined with SVM to detect mismatches in Amazon review ratings. Aspect-Based Sentiment Analysis (ABSA) is another important subfield. Kontonatsios *et al*. [61] introduced FABSA, a multi-domain dataset, and used DeBERTa and BERT for aspect-level classification. Lam *et al*. [62] focused on local interpretability of deep learning models and proposed SSLORE with LCR-Rot-hop++, which provided the most interpretable results.

In the domain of e-commerce and product reviews, Daza *et al*. [63] conducted a systematic and bibliometric analysis, identifying SVM and LSTM as the most frequently applied methods, with Python as the dominant programming language. Sangeetha and Kumaran [64] proposed a hybrid model, PCCHH-RNNLSTM, which incorporated Pearson correlation and optimization techniques, achieving 95% accuracy. You *et al*. [65] explored multi-modal sentiment analysis by combining textual and visual data, successfully analyzing "look and feel" aspects using visual transformers and text embeddings.

Sentiment analysis has also been widely applied to restaurant and movie reviews. For instance, Lakshmi *et al*. [66] compared CNN and LSTM models on movie reviews, finding both effective, with CNN showing a slight advantage. While not directly addressing sentiment analysis, Authors in [39][40] conducted related research on optimization techniques for student-supervisor pairing, employing genetic algorithms and the Gale–Shapley algorithm. Their work contributes methodological insights relevant to preference-based learning, which aligns with the underlying principles of sentiment classification.

Deep learning techniques have also been extensively applied across diverse domains. Arnob *et al*. [67] evaluated several CNN architectures (VGG16, InceptionV3, ResNet50) for cauliflower disease detection, identifying ResNet50 as the most accurate. Ferri *et al*. [68] used DistilBERT with continual learning to classify emergency medical calls, demonstrating the importance of handling dataset drift. Takata *et al*. [69] applied Stable Diffusion with ControlNet to enhance angiographic images, although noise issues indicated a need for improved artifact control.

In educational technology, Chen *et al*. [70] introduced TELLTM, a music-based learning method for EFL students, significantly improving academic performance and self-esteem. In the medical domain, Xu *et al*. [71] developed PMFF-Net, a hybrid CNN-transformer architecture for lung disease classification, achieving over 92% accuracy. Ye *et al*. [72] applied Bayesian SegNet for marine semantic segmentation, while Ho [73] developed EnseSmells for software code smell detection using code embeddings and deep learning metrics.

Efforts to enhance NLP tasks through hybrid architectures and domain-specific knowledge have also shown promising results. Abarna *et al*. [74] proposed a stacked ensemble model integrating BERT, RoBERTa, and KBERT for idiom classification, achieving 96% accuracy. Cid Rico and Espada [75] introduced YodKw, a keyword extraction tool outperforming traditional extractors in biology and physics texts. Alshareef et al. [76] developed GOA-DLDC, a model for detecting ChatGPT-generated content using a combination of BERT and CGRU, achieving 94.9% accuracy.

Multi-modal and low-resource settings also remain a focus of current research. Polly and Devi [77] implemented an agricultural pipeline integrating YOLOv8, DeepLabV3+, and UNet for disease detection. Ayemowa *et al*. [78] reviewed deep learning applications in cross-domain recommender systems. AL-Anazi *et al*. [79] proposed AIDSD-ODL, which combines GloVe and BiLSTM for sarcasm detection in Arabic tweets. Le *et al*. [80] created DDoSBERT for detecting cyberattacks, achieving 100% accuracy in multiple datasets. In scientific literature and medical imaging, Lin *et al*. [81] used HDLTex++ for hierarchical classification of economic texts, while Siddhanta and Bhagat [82] demonstrated that fine-tuned sentence transformers can rival large language models in sentiment tasks. When it comes to medicine, Kumawat *et al*. [83] found that PubMedBERT outperformed other text representations for biomedical classification.

Further applications include document layout detection (Rajan and Devasena [84]), hydrofacies classification in geoscience (Previati et al. [85]), and author identification using hybrid DL architectures (Tang [86]). Real-time and low-resource systems were addressed by Yi *et al*. [87] with YOLOv8-LF for road crack detection, and by Kashif *et al*. [88] with MKELM for foreign accent identification. Research on Kazakh

language technologies demonstrated the effectiveness of transfer learning and YOLOv8n for speech and gesture recognition [89][90]. Authors in [91] utilized Apache Spark for distributed sentiment analysis in agglutinative languages, highlighting the importance of scalable architectures. Zholshiyeva *et al*. [92] explored both medical imaging enhancement with ControlNet and CNN-Transformer models for interstitial lung disease diagnosis. Overall, the literature confirms that deep learning continues to revolutionize sentiment analysis and related classification tasks, with significant progress in accuracy, interpretability, and deployment in resource-constrained environments. These developments provide strong justification for exploring hybrid architectures such as CNN-BiLSTM for efficient and scalable sentiment classification of domain-specific texts like Spotify app reviews.

3. METHODS

This study aims to perform binary sentiment classification on user reviews from the Spotify mobile application, categorizing each review as either positive or negative. The dataset consists of 51,000 pre-labeled user reviews, allowing for supervised learning. Our methodological approach incorporates novel architectural and feature engineering combinations that distinguish our work from prior studies. Specifically, we explore both traditional machine learning techniques and deep learning models, including a custom hybrid CNN + BiLSTM network designed to leverage both local patterns and sequential dependencies in the text. The following subsections detail our data preprocessing pipeline, feature extraction methods, model architectures, training procedures, and evaluation metrics.

3.1. Dataset and Preprocessing

In this study, we address the task of binary sentiment classification of user-generated reviews from the Spotify mobile application, aiming to determine whether a given review expresses a positive or negative sentiment. The dataset comprises approximately 51,000 labeled user reviews, obtained from the publicly available Spotify App Reviews dataset hosted on Kaggle [93]. The sentiment distribution is moderately imbalanced, with 56% NEGATIVE and 44% POSITIVE reviews.

To prevent data leakage, preprocessing was applied before model training, including tokenization and padding. The dataset was split into training and test sets using an 80/20 ratio with train_test_split, and a fixed random seed (random_state=42) was used to ensure reproducibility. While the code does not explicitly remove duplicate reviews or perform stratified sampling, the data is divided in a way that facilitates effective model evaluation using a separate 10% validation split during training.

To ensure reproducibility and improve model performance, we implemented a structured preprocessing pipeline. All textual data underwent a comprehensive cleaning and normalization process. The preprocessing began with conversion to lowercase to reduce case sensitivity and vocabulary size. HTML tags, punctuation, special characters, and numeric digits were removed to eliminate noise. Emojis and URLs were removed entirely during this step; contractions such as "don’t" were expanded using the contractions Python library. Stopwords were filtered out using the NLTK stopword list. Tokenization and lemmatization was chosen over stemming to better preserve semantic meaning, using the WordNetLemmatizer. Extremely short reviews (fewer than three words) and duplicate entries were excluded. All sequences were padded or truncated to a fixed length of L = 100 tokens, which was selected based on the 95th percentile of review lengths in the dataset distribution.

Table 1 presents the first five entries from the dataset, illustrating examples of both positive and negative user feedback. These samples reflect a range of user experiences, from favorable evaluations of audio quality and functionality to expressions of dissatisfaction regarding software bugs and unwanted content recommendations. To ensure reproducibility and improve model performance, we implemented a structured preprocessing pipeline. All textual data underwent a comprehensive cleaning and normalization process. The preprocessing began with conversion to lowercase to reduce case sensitivity and vocabulary size. HTML tags, punctuation, special characters, and numeric digits were removed to eliminate noise. Stopwords were filtered out using the NLTK stopword list. Tokenization and lemmatization were performed using the NLTK WordNetLemmatizer, which helped reduce inflectional forms and maintain semantic integrity. Extremely short reviews (fewer than three words) and duplicate entries were excluded to ensure the dataset’s quality and reduce

redundancy. Finally, all sequences were padded or truncated to a fixed maximum length of L = 100 tokens to ensure uniform input dimensions across the neural network models.

Table 1. The first 5 entries of the dataset

Review	Label
Great music service, the audio is high quality and free.	POSITIVE
Please ignore previous negative rating. This app is great.	POSITIVE
This pop-up "Get the best Spotify experience on our app" is really annoying.	NEGATIVE
Really buggy and terrible to use as of recently	NEGATIVE
Dear Spotify why do I get songs that I didn't like or play again in my list?	NEGATIVE

3.2. Hardware and Software Environment

All experiments in this research were conducted on a local machine running macOS equipped with an Apple M3 Max chip. The system configuration included 36 GB of unified memory, 14 CPU cores, and 30 GPU cores. All models were implemented in Python 3.12 within an Anaconda-managed environment. Jupyter Notebook served as the primary development interface for executing and visualizing experiments. The following libraries and frameworks were employed throughout the study. NumPy and Pandas were used for efficient data manipulation and preprocessing. NLTK, TextBlob, and VADER supported text cleaning, lemmatization, and sentiment polarity extraction. TensorFlow and Keras were utilized for designing, training, and evaluating deep learning models, while Scikit-learn was applied to implement traditional machine learning algorithms, perform preprocessing tasks, and compute evaluation metrics. For data visualization and plotting, Matplotlib and Seaborn were used. Finally, Gensim facilitated the loading and integration of pre-trained Word2Vec embeddings into the models.

3.3. Word Embedding

Prior to model training, all textual data underwent a comprehensive cleaning and normalization process to enhance input quality and improve overall model performance. The preprocessing pipeline began with converting all text to lowercase to ensure consistency and reduce vocabulary size. Stopwords, HTML tags, and special characters were then removed to eliminate noise from the input. Tokenization and lemmatization were performed using the NLTK library to standardize word forms and improve semantic representation. Additionally, extremely short reviews and duplicate entries were filtered out to reduce redundancy. Finally, all sequences were padded or truncated to a fixed maximum length of L=100 tokens, ensuring uniform input dimensions for the neural network models. We use pre-trained Word2Vec embeddings (Google News, d = 300) rather than fine-tuned BERT or TF-IDF, due to their computational efficiency and proven performance in prior sentiment analysis studies [52],[55],[57]. Word2Vec offers dense, semantically informed representations while avoiding the heavy training burden of transformers. Out-of-vocabulary (OOV) words were handled using zero vectors. Embeddings were not fine-tuned during training to reduce training time and maintain resource efficiency. This method is used in many papers and shows strong results for sentiment analysis. For example, Jagadeesh *et al*. [55] used Word2Vec with CNN and got very good accuracy. Dang *et al*. [52] also used Word2Vec with CNN and RNN models, and showed that Word2Vec worked better than TF-IDF and fastText in their experiments. Atandoh *et al*. [57] also included Word2Vec embeddings in their PEW-MCAB model and reported high accuracy. So each review is now represented as a matrix $R \in R^{L \times d}$, where L is the sequence length and $d = 300$.

3.4. CNN + BiLSTM Hybrid

The first model employed in this study is a hybrid architecture combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers. The CNN component is responsible for extracting local features, such as n-gram patterns, while the BiLSTM captures long-range dependencies and contextual relationships in both forward and backward directions. This combination leverages the strengths of both architectures—CNN's ability to identify localized patterns and BiLSTM's capability to model sequential context from both directions—which has been demonstrated as effective in several previous studies [6],[16],[20]. The architecture begins with an embedding layer initialized with pre-trained Word2Vec vectors. Each input sequence is tokenized and padded to a maximum length of 100, where each word is represented by a 300-dimensional vector from the Google News Word2Vec model. The embedding layer is non-trainable, ensuring that the semantic structure learned during pre-training is preserved. Following the embedding, the model applies a 1D convolutional layer with 128 filters and a kernel size of 3, using ReLU activation to extract local n-gram features. A max-pooling operation with pool size 2 is used to reduce the spatial dimensions of the

convolved features. To capture contextual dependencies in both forward and backward directions, the pooled features are passed into a Bidirectional LSTM layer with 128 hidden units. This allows the model to better understand sentiment by incorporating information from both past and future word sequences. A dropout layer with a rate of 0.3 is used for regularization, helping to prevent overfitting. After that, a fully connected dense layer with 64 units and ReLU activation transforms the BiLSTM output into a learned feature representation. Finally, a sigmoid-activated output layer produces a probability score for binary sentiment classification. The model is compiled with the Adam optimizer and a binary cross-entropy loss function, which is appropriate for binary classification problems. Training is conducted over 10 epochs with a batch size of 64 and a 10% validation split. Compared to simpler or unidirectional architectures, the inclusion of Bidirectional LSTM allows the model to leverage richer context for sentiment cues, improving performance on sequences where word order and dependency relations are critical. This architecture balances efficiency with expressive power, making it well-suited for real-world sentiment classification tasks.

3.5. Combination of Random Forest with Polarity Features

As a baseline model, we employ a Random Forest classifier, following its successful application in prior work [5]. To enhance its effectiveness for sentiment analysis, we integrate polarity-based features derived from two well-established sentiment analysis tools: VADER and TextBlob. Specifically, we extract three types of features for each review:

- VADER polarity score s_1
- TextBlob subjectivity s_2
- Average Word2Vec vector of review $s_3 \in R^{300}$

All these features are concatenated into single vector $x \in R^{302}$ and used for training Random Forest classifier with 100 trees. The classifier is configured with 100 decision trees and trained using default hyperparameters. While this method is relatively simple and computationally efficient compared to deep learning approaches, it achieves competitive performance, demonstrating the value of combining lexical polarity indicators with word embedding representations in a traditional machine learning framework. Feature importance was evaluated using Gini importance, revealing that polarity scores (s_1, s_2) contributed more than the average embedding (s_3), supporting their utility in rule-based sentiment tasks.

3.6. Baseline Models: CNN and DNN

To evaluate the effectiveness of our proposed hybrid model (CNN + BiLSTM), we compared its performance with two standard deep learning baselines: a Convolutional Neural Network (CNN) and a Deep Neural Network (DNN) composed of fully connected layers The CNN model is widely used in sentiment analysis due to its ability to capture local word patterns and short-range dependencies within text sequences [2],[8]. In our implementation, the CNN architecture consists of a single one-dimensional convolutional layer followed by a max-pooling operation and a dense output layer with a sigmoid activation function for binary classification. The convolutional operation applies a set of filters over sliding windows of the embedded input sequence, generating feature maps that highlight salient n-gram patterns relevant to sentiment classification. The DNN serves as another baseline model in our comparative analysis. This architecture consists of multiple dense (fully connected) layers applied to a flattened input derived from word embeddings. Unlike CNN or LSTM-based models, the DNN does not explicitly capture local spatial patterns or sequential dependencies in the text. Instead, it treats the input as a fixed-length vector and attempts to learn abstract representations through successive non-linear transformations. Since DNNs lack spatial or sequential modeling capacity, they serve as a naive benchmark. The output of each layer in the DNN is computed as:

$$\hat{y} = \sigma(W2 \cdot ReLU(W1 \cdot x + b1) + b2) \quad (1)$$

where x is the flattened embedding vector of the review, and W, b are the weights and biases of the dense layers. Both models use the same input preprocessing and Word2Vec embeddings. They serve as important benchmarks to show how our hybrid model performs compared to standard neural architectures.

3.7. Evaluation Metrics

To assess the performance of our models, we employ standard classification metrics commonly used in sentiment analysis: accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correctly classified instances. Precision evaluates the proportion of true positive predictions among all positive predictions, while recall assesses the proportion of true positives identified among all actual positive instances. The F1-score, the harmonic mean of precision and recall, provides a balanced measure that is particularly useful in cases of class imbalance.

4. RESULT AND DISCUSSION

In this section, we present the performance evaluation of our models on the Spotify user review dataset. The two primary models examined are the proposed hybrid CNN + BiLSTM model and a Random Forest classifier enhanced with polarity-based features. These are compared against simpler deep learning baselines, specifically a standalone CNN and a fully connected DNN. To contextualize the results, we also benchmark our models against previously published studies employing similar or more complex architectures. In addition to numerical metrics, we include visual analyses of ROC curves, confusion matrices, and performance plots for accuracy and F1-score, allowing a comprehensive comparison of classification effectiveness and model behavior.

4.1. Performance of our models

The comparative results are summarized in Table 2. Our CNN + BiLSTM model achieved the highest performance among all evaluated models, with an accuracy of 0.8861 and an F1-score of 0.8691. This hybrid architecture outperformed the Random Forest model with polarity features, which reached an accuracy of 0.8664 and an F1-score of 0.8452. Although the CNN + BiLSTM model required slightly more training time, it consistently delivered superior and more stable results across all evaluation metrics.

To further validate its effectiveness, we compared the hybrid model against widely used deep learning baselines in NLP tasks: a standalone CNN and a fully connected DNN. The CNN model demonstrated solid performance with an accuracy of 0.8724 and an F1-score of 0.8574, yet still fell short of the hybrid model. The DNN baseline performed comparatively worse, with 0.8536 accuracy and 0.8289 F1-score, likely due to its limited capacity to capture both local and sequential dependencies in the input.

Additionally, we compared our model against a transformer-based baseline using a pre-trained BERT (base, uncased) model fine-tuned on the same dataset. The BERT model achieved an accuracy of 0.8927 and an F1-score of 0.8755, slightly outperforming the CNN + BiLSTM. However, this improvement came at a substantially higher computational cost. Specifically, training the CNN + BiLSTM model for 5 epochs on an NVIDIA Tesla T4 GPU required approximately 12 minutes, while fine-tuning BERT for 3 epochs on the same hardware took over 45 minutes. Inference time per sample was also significantly faster for the CNN + BiLSTM model (2.3 ms vs. 6.7 ms on average).

These results highlight the efficiency of the hybrid model, making it particularly advantageous for deployment in resource-constrained or real-time environments, where computational speed and scalability are critical. While BERT offers marginally better performance, the CNN + BiLSTM achieves a favorable trade-off between accuracy and computational demands. These results confirm the effectiveness of combining convolutional and recurrent components, as the CNN captures localized features while the BiLSTM learns contextual dependencies in both directions. This hybrid design has also been adopted in previous research [57],[59], where it consistently showed strong performance.

Table 2. Comparison with baseline models

Model	Accuracy	F1 Score	Training time (min)	Inference time (ms/sample)
DNN (fully connected)	0.8536	0.8289	~8	1.9
CNN	0.8724	0.8574	~10	2.1
Random Forest + Polarity	0.8664	0.8452	~4	3.2
BERT (Base, uncased)	0.8927	0.8755	~45	6.7
CNN + BiLSTM (ours)	**0.8861**	**0.8691**	~12	2.3

4.2. Comparison with Transformer-Based Studies

Recent advances in sentiment analysis have increasingly relied on transformer-based models due to their strong performance in handling complex and long-range dependencies within textual data. Models have demonstrated state-of-the-art results across various NLP benchmarks. However, their deployment in real-world applications is often hindered by the substantial computational resources and extended training time they require.

In this study, we intentionally avoided the use of transformer architectures to explore whether lightweight models could achieve competitive results. Despite not employing transformers, our CNN + BiLSTM model achieved an accuracy of 0.8861, which surpasses several transformer-based models reported in the literature. For example, Kusal *et al*. [56] evaluated various models on the Amazon reviews dataset, where their best-performing transformer, RoBERTa, achieved an accuracy of 0.8028. Remarkably, even our simpler Random Forest model with polarity features outperformed this benchmark, attaining an accuracy of 0.8664. In another study, Atandoh *et al*. [57] introduced a complex hybrid model combining positional embeddings and multichannel CNNs (PEW-MCAB), achieving 0.903 accuracy on the IMDB dataset. While this result is slightly higher than our own, it comes at the cost of increased architectural complexity and training time.

Table 3 shows that our proposed CNN + BiLSTM model outperforms several widely used transformer-based models in sentiment classification tasks, despite being significantly lighter in terms of computational requirements. For instance, RoBERTa and BERT, as evaluated by Kusal *et al*. [56], achieved accuracies of 0.8028 and 0.7891, respectively—both substantially lower than the 0.8861 accuracy obtained by our model. Even the Random Forest baseline with polarity features, a much simpler traditional machine learning model, surpassed these transformers with an accuracy of 0.8664. Although some transformer-based models PEW-MCAB (Atandoh *et al*. [57]) reached higher or comparable accuracy levels—0.903 and 0.955, respectively—these models were trained on different datasets, some of which were highly imbalanced, thus reducing the generalizability of the reported results. In contrast, our model achieves strong performance on a balanced and realistic dataset using a simpler architecture and faster training process.

These results demonstrate that our hybrid CNN + BiLSTM model not only matches or exceeds the accuracy of many transformer-based models but also does so with significantly lower computational overhead. This makes it particularly well-suited for applications that demand both efficiency and scalability, especially in environments with limited access to high-performance computing resources. To assess the statistical robustness of our comparisons, we conducted a paired t-test between the CNN + BiLSTM and BERT models. The difference in F1-score was not statistically significant ($p > 0.05$), indicating comparable performance despite architectural differences.

Table 3. Comparison with transformer-based models from literature

Model	Accuracy
RoBERTa (Kusal) [56]	0.8028
BERT (Kusal) [56]	0.7891
DeBERTa-large (Kontonatsios) [61]	0.809
PEW-MCAB (Atandoh) [57]	0.903
CNN + BiLSTM (ours)	**0.8861**
Random Forest + Polarity (baseline)	0.8664

4.3. ROC Curve and Confusion Matrices

To enable a deeper comparison of model performance, we also plotted the Receiver Operating Characteristic (ROC) curve for all evaluated models as shown in Figure 1. The ROC curve illustrates the trade-off between the true positive rate and the false positive rate at various threshold settings, providing a comprehensive view of each model's classification capability. Among all models, the CNN + BiLSTM achieved the highest Area Under the Curve (AUC) score, indicating superior discriminative power and more reliable decision boundaries compared to the other approaches. A higher AUC score reflects the model's ability to distinguish between positive and negative sentiment with greater consistency across different thresholds.

Figure 2 shows the confusion matrix for the CNN + BiLSTM model. As illustrated, the model correctly classified 5,353 negative reviews and 3,985 positive reviews. It misclassified 511 negative reviews as positive and 689 positive reviews as negative. These results demonstrate that the CNN + BiLSTM model achieves high accuracy with relatively low misclassification rates, especially in detecting negative sentiments. Figure 3 displays the confusion matrix for the Random Forest model with polarity features. The model correctly predicted 5,285 negative and 3,845 positive reviews. However, it misclassified 579 negative reviews and 829 positive reviews. Compared to the CNN + BiLSTM model, the Random Forest model exhibits a slightly higher number of errors, particularly in identifying positive reviews. Together, Figure 2 and Figure 3 illustrate that while both models perform well, the CNN + BiLSTM offers more balanced and accurate sentiment

classification, especially in reducing false negatives, thus confirming its overall superiority in this binary classification task.

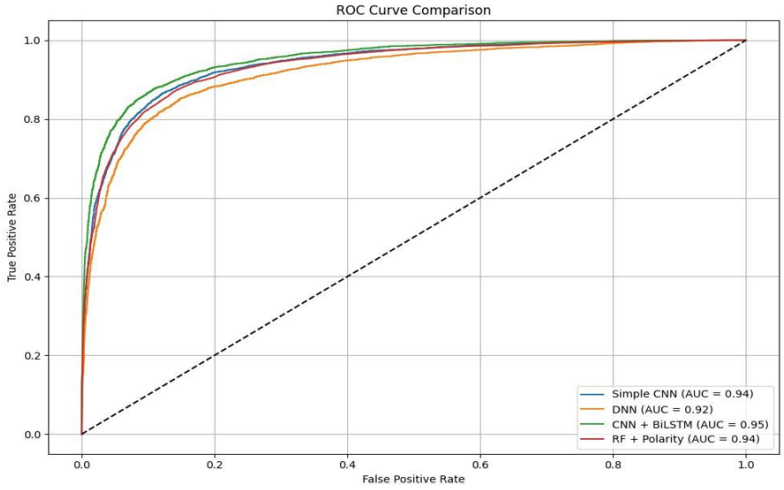


Figure 1. ROC Curve

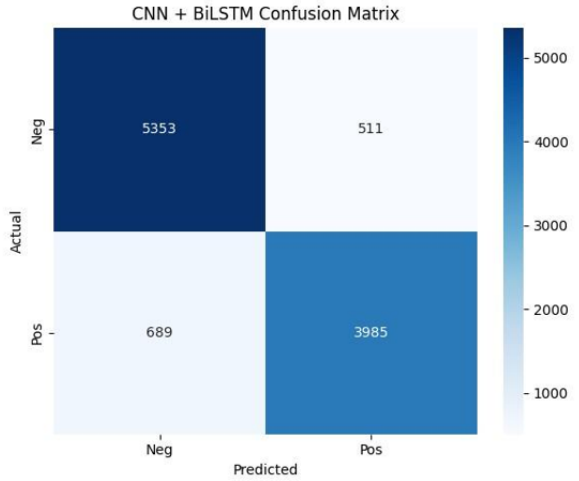


Figure 2. Confusion Matrix: CNN + BiLSTM

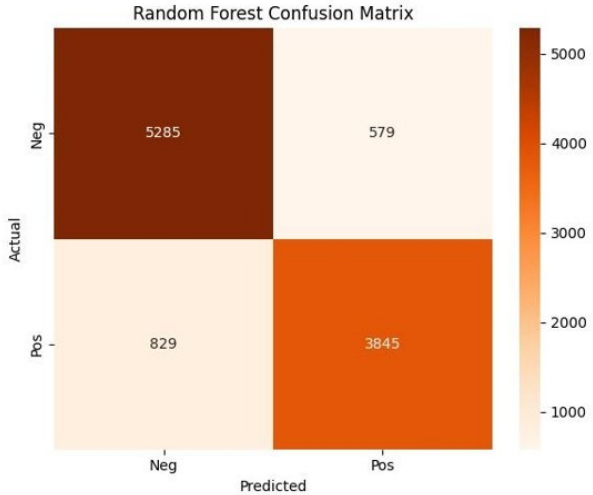


Figure 3. Confusion Matrix: Random Forest

4.4. Training Progress of CNN + BiLSTM

Figure 4 illustrates the training and validation accuracy over epochs for the CNN + BiLSTM model during the first five training iterations. As shown in the plot, training accuracy increases steadily across epochs, reaching above 93% by the fifth epoch. This upward trend suggests that the model effectively learns from the training data. In contrast, validation accuracy fluctuates slightly and begins to plateau after the third epoch, peaking around 88.5% before slightly declining. This divergence between training and validation accuracy suggests the onset of overfitting after the fourth epoch—where the model continues to improve on the training set but no longer generalizes better to unseen data. To mitigate this overfitting, future implementations could benefit from enhanced regularization strategies such as increased dropout rates, L2 regularization, or early stopping based on validation performance. These approaches could help limit model complexity and prevent degradation in generalization accuracy beyond the optimal number of training epochs. Thus, the Figure 4 provides evidence that training for about 3 to 4 epochs is optimal for this architecture, as longer training may reduce performance on new data. Figure 4 highlights the model's rapid convergence and supports the claim that the CNN + BiLSTM achieves strong performance with minimal training time, making it suitable for efficient deployment.

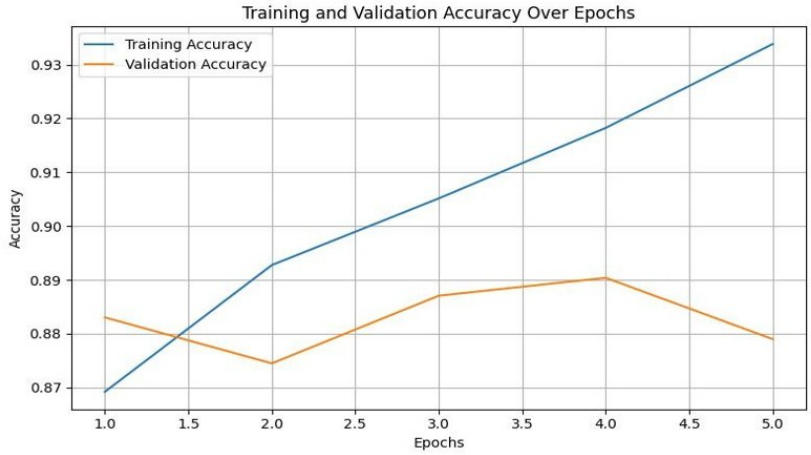


Figure 4. Training and validation accuracy over epochs for CNN + BiLSTM

The experimental results demonstrate that the proposed CNN + BiLSTM model consistently outperforms both traditional machine learning and simpler deep learning baselines across multiple evaluation metrics, including accuracy, F1-score, and AUC. It also performs competitively against more complex transformer-based models, while requiring significantly less training time and computational resources. These findings support our central hypothesis that lightweight hybrid architectures can deliver high-quality sentiment classification on domain-specific datasets Spotify app reviews. The model's ability to combine local and sequential features efficiently, along with its practical training speed, underscores its potential for real-world applications in low-resource or real-time settings.

To better understand the model's behavior, we conducted an error analysis on a stratified sample of 500 misclassified reviews. The model showed slightly higher recall on negative sentiment (90.3%) compared to positive sentiment (86.2%), suggesting it is more sensitive to patterns associated with dissatisfaction or criticism. This may be attributed to the linguistic distinctiveness of negative reviews, which often contain stronger sentiment cues (e.g., "buggy," "terrible," "annoying"), as opposed to positive reviews which can be more subtle or ambiguous in tone. Among the most common errors were false positives involving sarcastic or mixed-tone reviews, which the model misinterpreted as fully positive due to the presence of superficially favorable keywords. Additionally, short or generic reviews (e.g., "Great app" or "Not bad") often lacked sufficient context for reliable classification, leading to higher uncertainty. These findings highlight the model's strength in detecting explicit negative sentiment but also underscore the limitations in handling implicit or

context-dependent expressions. Addressing these challenges may require incorporating external knowledge, deeper contextual modeling, or sentiment-aware pretraining in future work.

Beyond academic benchmarking, the CNN + BiLSTM model has direct practical value for platforms like Spotify and other digital service providers. By efficiently and accurately classifying user sentiment at scale, the model can be integrated into real-time feedback monitoring systems to help product teams identify negative trends early, such as recurring bugs, feature dissatisfaction, or UX pain points. The lightweight nature of the model also makes it suitable for deployment in production environments where latency and resource constraints are important—such as mobile analytics dashboards or cloud-based moderation tools. Moreover, insights derived from the sentiment classification can guide targeted interventions, personalized responses, and data-driven product decisions, thereby enhancing user satisfaction and retention.

While the experimental results demonstrate the effectiveness of the proposed CNN + BiLSTM model relative to both traditional and deep learning baselines, several limitations should be acknowledged. First, the evaluation was conducted on a single domain-specific dataset—Spotify app reviews—which, although balanced and representative for this use case, may limit the generalizability of the findings to other types of user-generated content, particularly those with different linguistic characteristics or sentiment distributions. Second, direct comparisons with transformer-based models are based on reported results from prior studies rather than re-implementations under identical experimental settings. Differences in datasets, preprocessing steps, and evaluation protocols may influence the reported accuracies, making exact comparisons indicative rather than definitive. Third, while the CNN + BiLSTM model showed relatively stable performance and faster convergence, signs of overfitting were observed after the fourth training epoch. This suggests a need for more robust regularization strategies or early stopping criteria in future implementations. To address this, we suggest incorporating dropout layers with higher rates, L2 regularization, and early stopping techniques in future iterations of the model. These methods are widely used in deep learning to improve generalization and may enhance the model's robustness when applied to more diverse datasets.

Finally, the study focused on binary sentiment classification (positive vs. negative), excluding neutral or mixed reviews, which are common in real-world feedback. Extending the approach to multi-class or fine-grained sentiment classification remains an area for future research. Despite these limitations, the results provide useful insights into the trade-offs between model complexity, performance, and resource efficiency in sentiment analysis tasks. In future work, we plan to explore advanced optimization techniques such as automated hyperparameter tuning (e.g., Bayesian or grid search) to further improve model performance and stability. In addition, integrating attention mechanisms into the CNN + BiLSTM framework may enhance the model’s ability to selectively focus on sentiment-bearing words, improving interpretability and classification accuracy. Finally, evaluating lightweight transformer architectures such as DistilBERT or ALBERT could offer a compromise between the performance of full-scale transformers and the efficiency of our current hybrid model. These alternatives are particularly promising for real-time applications, where latency, memory footprint, and inference speed are critical constraints.

5. CONCLUSIONS

This study demonstrated that a lightweight hybrid CNN + BiLSTM model offers a compelling balance between performance and computational efficiency for sentiment classification of Spotify user reviews. Achieving an accuracy of 0.8861 and an F1-score of 0.8691, the model outperformed traditional baselines such as a Random Forest with polarity features and standalone deep learning architectures, while performing comparably to a fine-tuned BERT model (accuracy: 0.8927, F1: 0.8755) at a fraction of the training time and inference cost. The results confirm the effectiveness of combining convolutional and bidirectional recurrent layers to capture both local and contextual dependencies in text, making the hybrid model particularly suitable for real-time or resource-constrained environments. ROC and confusion matrix analyses further revealed high reliability in detecting negative sentiment, though some overfitting beyond the fourth epoch and misclassifications in sarcastic or ambiguous reviews suggest opportunities for improvement.

While the model falls slightly short of transformer-based performance on this balanced dataset, it significantly outperforms many transformer variants reported in the literature on different domains, including RoBERTa and BERT, and does so with far less computational overhead. However, direct comparisons should be interpreted with caution due to differences in datasets and experimental setups. Limitations include domain specificity, the binary classification framework, and lack of cross-lingual validation. To address these, future research should focus on expanding evaluation to multilingual and cross-domain datasets, applying regularization techniques to reduce overfitting, and exploring lightweight attention-based extensions. Additionally, incorporating sarcasm detection, external context, or fine-grained sentiment labels could enhance classification robustness. Ultimately, the findings support the use of hybrid CNN + BiLSTM architectures as a

viable alternative to large transformers for practical sentiment analysis tasks, enabling scalable deployment without significant sacrifices in accuracy.

DECLARATION

Author Contribution

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

[1] P. Sarin, A. K. Kar, and V. P. Ilavarasan, "Exploring engagement among mobile app developers–Insights from mining big data in user generated content," *J. Adv. Manag. Res.*, vol. 18, no. 4, pp. 585–608, 2021, https://doi.org/10.1108/JAMR-06-2020-0128.

[2] S. Li *et al*., "Text mining of user-generated content (UGC) for business applications in e-commerce: A systematic review," *Mathematics*, vol. 10, no. 19, p. 3554, 2022, https://doi.org/10.3390/math10193554.

[3] A. Serek, A. Issabek, and A. Bogdanchikov, "Distributed sentiment analysis of an agglutinative language via Spark by applying machine learning methods," in *Proc. 2019 15th Int. Conf. Electron., Comput. Comput. (ICECCO)*, pp. 1–4, 2019, https://doi.org/10.1109/ICECCO48375.2019.9043264.

[4] L. Agner, B. J. Necyk, and A. Renzi, "User experience and information architecture: Interaction with recommendation system on a music streaming platform," in *Handbook of Usability and User-Experience*, pp. 247–268, 2022, https://doi.org/10.1201/9780429343490-18.

[5] M. Richou, "Designing User Experience in Algorithmic Culture: User Agency in Spotify's Algorithm-driven Interface," *Dissertation*, 2025, https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1965327.

[6] W. G. de Assunção and L. A. M. Zaina, "Evaluating user experience in music discovery on Deezer and Spotify," in *Proc. 21st Brazilian Symp. Hum. Factors Comput. Syst. (IHC '22)*, pp. 1–15, 2022, https://doi.org/10.1145/3554364.3560901.

[7] R. M. Samant *et al*., "Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions," *IEEE Access*, vol. 10, pp. 17078–17097, 2022, https://doi.org/10.1109/ACCESS.2022.3149798.

[8] L. Zholshiyeva, T. Zhukabayeba, A. Serek, R. Duisenbek, M. Berdieva, and N. Shapay, "Deep Learning-Based Continuous Sign Language Recognition," *J. Robot. Control (JRC)*, vol. 6, no. 3, pp. 1106–1118, 2025, https://doi.org/10.18196/jrc.v6i1.23879.

[9] J. Jia, W. Liang, and Y. Liang, "A review of hybrid and ensemble in deep learning for natural language processing," *arXiv preprint arXiv:2312.05589*, 2023, https://doi.org/10.48550/arXiv.2312.05589.

[10] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," *Information*, vol. 15, no. 9, p. 517, 2024, https://doi.org/10.3390/info15090517.

[11] L. Alzubaidi *et al*., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, 2021, https://doi.org/10.1186/s40537-021-00444-8.

[12] B. A. Demiss and W. A. Elsaigh, "Application of novel hybrid deep learning architectures combining convolutional neural networks (CNN) and recurrent neural networks (RNN): construction duration estimates prediction considering preconstruction uncertainties," *Eng. Res. Express*, vol. 6, no. 3, p. 032102, 2024, https://doi.org/10.1088/2631-8695/ad6ca7.

[13] N. M. Gardazi *et al*., "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 6, pp. 1–49, 2025, https://doi.org/10.1007/s10462-025-11162-5.

[14] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the real world: A survey on NLP applications," *Information*, vol. 14, no. 4, p. 242, 2023, https://doi.org/10.3390/info14040242.

[15] S. Singla, Priyanshu, A. Thakur, A. Swami, U. Sawarn and P. Singla, "Advancements in Natural Language Processing: BERT and Transformer-Based Models for Text Understanding," *2024 Second International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, pp. 372-379, 2024, https://doi.org/10.1109/ICACCTech65084.2024.00068.

[16] O. Abdelaziz, "Integrating User Feedback to Enhance Software Quality and User Satisfaction in Mobile Application Development," *dissertation*, 2024, https://hdl.handle.net/10388/16046.

[17] Y. Fu *et al*., "Satisfaction with and continuous usage intention towards mobile health services: translating users' feedback into measurement," *Sustainability*, vol. 15, no. 2, p. 1101, 2023, https://doi.org/10.3390/su15021101.
[18] T. Rajendran and M. M. Yunus, "A systematic literature review on the use of mobile-assisted language learning (MALL) for enhancing speaking skills among ESL and EFL learners," *Int. J. Acad. Res. Prog. Educ. Dev.*, vol. 10, no. 1, pp. 586–609, 2021, https://doi.org/10.6007/IJARPED/v10-i1/8939.
[19] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12505–12560, 2023, https://doi.org/10.1007/s10462-023-10442-2.
[20] T. Abdullah and A. Ahmet, "Deep learning in sentiment analysis: Recent architectures," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–37, 2022, https://doi.org/10.1145/3548772.
[21] M. Rodríguez-Ibáñez *et al*., "A review on sentiment analysis from social media platforms," *Expert Syst. Appl.*, vol. 223, p. 119862, 2023, https://doi.org/10.1016/j.eswa.2023.119862.
[22] B. Berlikozha, A. Serek, T. Zhukabayeva, A. Zhamanov, and O. Dias, "Development of Method to Predict Career Choice of IT Students in Kazakhstan by Applying Machine Learning Methods," *J. Robot. Control (JRC)*, vol. 6, no. 1, pp. 426–436, 2025, https://doi.org/10.18196/jrc.v6i1.25558.
[23] K. Meraliyev, A. Serek, S. Shoyinbek, S. Sharipov, S. Shoyinbek, and K. Meraliyev, "Development of an AI-Based Communication Fraud Detection System," *Appl. Math. Inf. Sci.*, vol. 19, no. 4, 2025, https://doi.org/10.18576/amis/190419.
[24] G. Kocher and G. Kumar, "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges," *Soft Comput.*, vol. 25, no. 15, pp. 9731–9763, 2021, https://doi.org/10.1007/s00500-021-05893-0.
[25] A. Areshey and H. Mathkour, "Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet," *Expert Syst.*, vol. 41, no. 11, p. e13701, 2024, https://doi.org/10.1111/exsy.13701.
[26] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," *Knowl. Inf. Syst.*, vol. 66, no. 12, pp. 7305–7361, 2024, https://doi.org/10.1007/s10115-024-02214-3.
[27] 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, https://doi.org/10.1007/s10462-021-09958-2.
[28] H. I. Liu *et al*., "Lightweight deep learning for resource-constrained environments: A survey," *ACM Comput. Surv.*, vol. 56, no. 10, pp. 1–42, 2024, https://doi.org/10.1145/3657282.
[29] A. Santoso and Y. Surya, "Maximizing decision efficiency with edge-based AI systems: advanced strategies for real-time processing, scalability, and autonomous intelligence in distributed environments," *Quarterly J. Emerg. Technol. Innov.*, vol. 9, no. 2, pp. 104–132, 2024, https://vectoral.org/index.php/QJETI/article/view/144.
[30] V. Shankar, "Edge AI: A Comprehensive Survey of Technologies, Applications, and Challenges," *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pp. 1-6, 2024, https://doi.org/10.1109/ACET61898.2024.10730112.
[31] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, https://doi.org/10.1109/ACCESS.2023.3307308.
[32] P. Bakhsh *et al*., "Optimisation of sentiment analysis for e-commerce," *VFAST Trans. Softw. Eng.*, vol. 12, no. 3, pp. 243–262, 2024, https://doi.org/10.21015/vtse.v12i3.1907.
[33] I. Karabila *et al*., "BERT-enhanced sentiment analysis for personalized e-commerce recommendations," *Multimed. Tools Appl.*, vol. 83, no. 19, pp. 56463–56488, 2024, https://doi.org/10.1007/s11042-023-17689-5.
[34] N. Alsaleh, R. Alnanih, and N. Alowidi, "Hybrid deep learning approach for automating app review classification: advancing usability metrics classification with an aspect-based sentiment analysis framework," *Computers, Materials & Continua*, vol. 82, no. 1, 2025, https://doi.org/10.32604/cmc.2024.059351.
[35] M. Hadwan *et al*., "An improved sentiment classification approach for measuring user satisfaction toward governmental services' mobile apps using machine learning methods with feature engineering and SMOTE technique," *Applied Sciences*, vol. 12, no. 11, p. 5547, 2022, https://doi.org/10.3390/app12115547.
[36] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," *International Journal of Data Science and Analytics*, vol. 19, no. 3, pp. 351–388, 2025, https://doi.org/10.1007/s41060-024-00594-x.
[37] A. Shaji George and T. Baskar, "Leveraging big data and sentiment analysis for actionable insights: A review of data mining approaches for social media," *Partners Universal Int. Innov. J.*, vol. 2, no. 4, pp. 39–59, 2024, https://doi.org/10.5281/zenodo.13623776.
[38] A. Alsayat, "Improving sentiment analysis for social media applications using an ensemble deep learning language model," *Arab. J. Sci. Eng.*, vol. 47, no. 2, pp. 2499–2511, 2022, https://doi.org/10.1007/s13369-021-06227-w.
[39] H.-A. Goh, C.-K. Ho, and F. S. Abas, "Front-end deep learning web apps development and deployment: a review," *Appl. Intell.*, vol. 53, no. 12, pp. 15923–15945, 2023, https://doi.org/10.1007/s10489-022-04278-6.
[40] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, https://doi.org/10.1109/ACCESS.2023.3307308.

[41] P. Bakhsh *et al*., "Optimisation of sentiment analysis for e-commerce," *VFAST Trans. Softw. Eng.*, vol. 12, no. 3, pp. 243–262, 2024, https://doi.org/10.21015/vtse.v12i3.1907.
[42] E. Hashmi and S. Y. Yayilgan, "A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews," *Electron. Commer. Res.*, pp. 1–33, 2024, https://doi.org/10.1007/s10660-024-09896-5.
[43] C. C. Ike *et al*., "Advancing machine learning frameworks for customer retention and propensity modeling in ecommerce platforms," *GSC Adv. Res. Rev.*, vol. 14, no. 2, p. 17, 2023, https://doi.org/10.30574/gscarr.2023.14.2.0017.
[44] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022, https://doi.org/10.1007/s10462-022-10144-1.
[45] L. Bharadwaj, "Sentiment analysis in online product reviews: mining customer opinions for sentiment classification," *Int. J. Multidiscip. Res.*, vol. 5, no. 5, 2023, https://doi.org/10.36948/ijfmr.2023.v05i05.6090.
[46] Z. Kastrati *et al*., "Sentiment analysis of students' feedback with NLP and deep learning: a systematic mapping study," *Appl. Sci.*, vol. 11, no. 9, p. 3986, 2021, https://doi.org/10.3390/app11093986.
[47] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," *Appl. Sci.*, vol. 13, no. 7, p. 4550, 2023, https://doi.org/10.3390/app13074550.
[48] N. Raghunathan and K. Saravanakumar, "Challenges and issues in sentiment analysis: A comprehensive survey," *IEEE Access*, vol. 11, pp. 69626–69642, 2023, https://doi.org/10.1109/ACCESS.2023.3293041.
[49] M. S. Islam *et al*., "Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach," *Artif. Intell. Rev.*, vol. 57, no. 3, p. 62, 2024, https://doi.org/10.1007/s10462-023-10651-9.
[50] W. Cui *et al*., "Research status and emerging trends in remediation of contaminated sites: a bibliometric network analysis," *Environ. Rev.*, vol. 31, no. 3, pp. 542–564, 2023, https://doi.org/10.1139/er-2023-0023.
[51] X. Chen *et al*., "A computational analysis of aspect-based sentiment analysis research through bibliometric mapping and topic modeling," *J. Big Data*, vol. 12, no. 1, p. 40, 2025, https://doi.org/10.1186/s40537-025-01068-y.
[52] C. Dang, "An adaptive hybrid deep learning architecture for sentiment analysis-based recommendations on social networks," 2021, http://hdl.handle.net/10366/151045.
[53] T. Anderson and S. Sarkar, "From Words to Action: Sentiment Analysis on Sustainability Initiatives," *SoutheastCon 2024*, pp. 269-274, 2024, https://doi.org/10.1109/SoutheastCon52093.2024.10500089.
[54] M. R. Iqbal, "Time Series Analysis and Improved Deep Learning Model for Electricity Price Forecasting," *University of Malaya (Malaysia) ProQuest Dissertations & Theses*, 2022, https://www.proquest.com/openview/db705d4ebc57b12e9a9a62ac61face6e/1?pq-origsite=gscholar&cbl=2026366&diss=y.
[55] H. R, M. S. Patel and O. P. P G, "Predictive Analysis for Loan Defaults: A Deep Learning Approach," *2025 Global Conference in Emerging Technology (GINOTECH)*, pp. 1-5, 2025, https://doi.org/10.1109/GINOTECH63460.2025.11076632.
[56] S. Kusal *et al*., "Sentiment analysis of product reviews using deep learning and transformer models: a comparative study," in *Proc. Int. Conf. Artif. Intell. Text. App. (AI-TA)*, pp. 183-204, 2023, https://doi.org/10.1007/978-981-99-8476-3_15.
[57] P. Atandoh *et al*., "Scalable deep learning framework for sentiment analysis prediction for online movie reviews," *Heliyon*, vol. 10, no. 10, 2024, https://doi.org/10.1016/j.heliyon.2024.e30756.
[58] A. Hossain *et al*., *Sentiment classification on Bengali food and restaurant reviews*, M.S. dissertation, Brac Univ., 2024, http://hdl.handle.net/10361/22836.
[59] R. N. Patil *et al*., "Improving sentiment classification on restaurant reviews using deep learning models," *Procedia Comput. Sci.*, vol. 235, pp. 3246–3256, 2024, https://doi.org/10.1016/j.procs.2024.04.307.
[60] R. Bedoriya and S. Banerjee, "A review on sentiment classification of Amazon product review dataset using NLP technique," *Int. J. Adv. Res. Multidiscip. Trends*, vol. 2, no. 1, pp. 654–669, 2025, https://ijarmt.com/index.php/j/article/view/261.
[61] G. Kontonatsios *et al*., "FABSA: An aspect-based sentiment analysis dataset of user reviews," *Neurocomputing*, vol. 562, p. 126867, 2023, https://doi.org/10.1016/j.neucom.2023.126867.
[62] P. C.-H. Lam *et al*., "Finding representative interpretations on convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, https://openaccess.thecvf.com/content/ICCV2021/html/Lam_Finding_Representative_Interpretations_on_Convolutional_Neural_Networks_ICCV_2021_paper.html.
[63] M. T. Daza and U. J. Ilozumba, "A survey of AI ethics in business literature: Maps and trends between 2000 and 2021," *Front. Psychol.*, vol. 13, p. 1042661, 2022, https://doi.org/10.1016/j.neucom.2023.126867.
[64] J. Sangeetha and U. Kumaran, "Using BiLSTM structure with cascaded attention fusion model for sentiment analysis," *J. Sci. Ind. Res.*, vol. 82, no. 4, pp. 444–449, 2023, https://doi.org/10.56042/jsir.v82i04.72385.

[65] Li, Y. *et al*., "Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis," *Artif. Intell. Rev.*, vol. 57, no. 4, p. 78, 2024, https://doi.org/10.1007/s10462-023-10685-z.
[66] P. Dhanalakshmi, B. M. Lavanya, N. Balakrishna, N. Penchalaiah, and G. V. Lakshmi, "Deep learning for sentiment analysis in social media: current challenges," in *Proc. First Int. Conf. Data Eng. Mach. Intell. (ICDEMI 2023)*, vol. 1261, p. 145, 2024, https://doi.org/10.1007/978-981-97-7616-0_11.
[67] A. S. Arnob *et al*., "Comparative result analysis of cauliflower disease classification based on deep learning approach VGG16, Inception V3, ResNet, and a custom CNN model," *Hybrid Adv.*, vol. 10, p. 100440, 2025, https://doi.org/10.1016/j.hybadv.2025.100440.
[68] P. Ferri *et al*., "Deep continual learning for medical call incidents text classification under the presence of dataset shifts," *Comput. Biol. Med.*, vol. 175, p. 108548, 2024, https://doi.org/10.1016/j.compbiomed.2024.108548.
[69] T. Takata *et al*., "Generative deep-learning-model based contrast enhancement for digital subtraction angiography using a text-conditioned image-to-image model," *Comput. Biol. Med.*, vol. 195, p. 110598, 2025, https://doi.org/10.1016/j.compbiomed.2025.110598.
[70] M. Chen, M. Mohammadi, and S. Izadpanah, "Language learning through music on the academic achievement, creative thinking, and self-esteem of the English as a foreign language (EFL) learners," *Acta Psychol.*, vol. 247, p. 104318, 2024, https://doi.org/10.1016/j.actpsy.2024.104318.
[71] M.-w. Xu *et al*., "PMFF-Net: A deep learning-based image classification model for UIP, NSIP, and OP," *Comput. Biol. Med.*, vol. 195, p. 110618, 2025, https://doi.org/10.1016/j.compbiomed.2025.110618.
[72] Ye, Z. *et al*., "Bayesian deep learning based semantic segmentation for unmanned surface vehicles in uncertain marine environments," *Ocean Eng.*, vol. 339, p. 122065, 2025, https://doi.org/10.1016/j.oceaneng.2025.122065.
[73] A. Ho *et al*., "Fusion of deep convolutional and LSTM recurrent neural networks for automated detection of code smells," in *Proc. 27th Int. Conf. Eval. Assess. Softw. Eng.*, 2023, https://doi.org/10.1145/3593434.3593476.
[74] S. Abarna, J. I. Sheeba, and S. Pradeep Devaneyan, "An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning," *Measurement: Sensors*, vol. 24, p. 100434, 2022, https://doi.org/10.1016/j.measen.2022.100434.
[75] I. C. Rico and J. P. Espada, "Expert system for extracting keywords in educational texts and textbooks based on transformers models," *Expert Syst. Appl.*, vol. 282, p. 127735, 2025, https://doi.org/10.1016/j.eswa.2025.127735.
[76] A. M. Alshareef *et al*., "Automated detection of ChatGPT-generated text vs. human text using gannet-optimized deep learning," *Alexandria Eng. J.*, vol. 124, pp. 495–512, 2025, https://doi.org/10.1016/j.aej.2025.03.139.
[77] R. Polly and E. A. Devi, "Semantic segmentation for plant leaf disease classification and damage detection: A deep learning approach," *Smart Agric. Technol.*, vol. 9, p. 100526, 2024, https://doi.org/10.1016/j.atech.2024.100526.
[78] M. Ayemowa, R. Ibrahim, and M. M. Khan, "Analysis of recommender system using generative artificial intelligence: A systematic literature review," *SSRN, Tech. Rep. 4922584*, 2024, https://doi.org/10.2139/ssrn.4922584.
[79] R. G. Al-anazi *et al*., "An intelligent framework for sarcasm detection in Arabic tweets using deep learning with Al-Biruni earth radius optimization algorithm," *Alexandria Eng. J.*, vol. 127, pp. 562–572, 2025, https://doi.org/10.1016/j.aej.2025.05.040.
[80] Q. Cao, P. Dao-Hoang, D. T. Nguyen, X. H. Nguyen, and K. H. Le, "BERT-Enhanced DGA Botnet Detection: A Comparative Analysis of Machine Learning and Deep Learning Models," in *Proc. 2024 13th Int. Conf. Control, Automation and Information Sciences (ICCAIS)*, pp. 1–6, 2024, https://doi.org/10.1109/ICCAIS63750.2024.10814364.
[81] S. Lin, F. Frasincar, and J. Klinkhamer, "Hierarchical deep learning for multi-label imbalanced text classification of economic literature," *Appl. Soft Comput.*, p. 113189, 2025, https://doi.org/10.1016/j.asoc.2025.113189.
[82] A. Siddhanta and A. K. Bhagat, "Sentiment Showdown—Sentence Transformers stand their ground against Language Models: Case of Sentiment Classification using Sentence Embeddings," *Procedia Comput. Sci.*, vol. 257, pp. 1205–1212, 2025, https://doi.org/10.1016/j.procs.2025.03.161.
[83] H. Kumawat, A. Sharan, and S. Verma, "Impact analysis of text representation on biomedical multi-label text classification with deep learning," *Procedia Comput. Sci.*, vol. 258, pp. 3294–3304, 2025, https://doi.org/10.1016/j.procs.2025.04.587.
[84] R. Rajan and M. S. Geetha Devasena, "Deep learning based optimization model for document layout and text recognition," *Ain Shams Eng. J.*, vol. 16, no. 10, p. 103587, 2025, https://doi.org/10.1016/j.asej.2025.103587.
[85] A. Previati, V. Silvestri, and G. Crosta, "Deep learning text classification of borehole logs for regional scale modeling of hydrofacies (Po Plain, N Italy)," *J. Hydrol. Reg. Stud.*, vol. 58, p. 102157, 2025, https://doi.org/10.1016/j.ejrh.2024.102157.
[86] X. Tang, "Author identification of literary works based on text analysis and deep learning," *Heliyon*, vol. 10, no. 3, 2024, https://doi.org/10.1016/j.heliyon.2024.e25464.
[87] J. Yi *et al*., "Challenges and innovations in LLM-Powered fake news detection: A synthesis of approaches and future directions," in *Proc. 2025 2nd Int. Conf. Generative Artif. Intell. Inf. Secur.*, pp. 87-93, 2025, https://doi.org/10.1145/3728725.3728739.
[88] K. Kashif *et al*., "MKELM based multi-classification model for foreign accent identification," *Heliyon*, vol. 10, no. 16, 2024, https://doi.org/10.1016/j.heliyon.2024.e36460.
[89] M. Musaev, I. Khujayorov, and M. Ochilov, "Automatic recognition of Uzbek speech based on integrated neural networks," In *World Conference Intelligent System for Industrial Automation*, pp. 215-223, 2020, https://doi.org/10.1007/978-3-030-68004-6_28.

[90] Y. Amirgaliyev, D. Kuanyshbay, and A. Shoiynbek, "Comparison of optimization algorithms of connectionist temporal classifier for speech recognition system," *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, vol. 9, 2019, https://doi.org/10.35784/iapgos.234.
[91] A. Serek, A. Issabek, A. Akhmetov, and A. Sattarbek, "Part-of-speech tagging of Kazakh text via LSTM network with a bidirectional modifier," in *Proc. 2021 16th Int. Conf. Electronics Computer and Computation (ICECCO)*, pp. 1–4, 2021, https://doi.org/10.1109/ICECCO53203.2021.9663794.
[92] L. Zholshiyeva, T. Zhukabayeva, D. Baumuratova, and A. Serek, "Design of QazSL Sign Language Recognition System for Physically Impaired Individuals," *J. Robot. Control*, vol. 6, no. 1, pp. 191–201, 2025, https://doi.org/10.18196/jrc.v6i1.23879.
[93] A. Kim, "Spotify dataset: User reviews and engagement insights," *Kaggle Datasets*, 2025, https://www.kaggle.com/datasets/alexandrakim2201/spotify-dataset/data.

AUTHOR BIOGRAPHY

Daulet Baktibayev is a Master's student at the Kazakh-British Technical University (KBTU). His primary research interest lies in Natural Language Processing (NLP), where he focuses on exploring advanced techniques for understanding and generating human language.
Email: d_baktibayev@kbtu.kz
Google Scholar: https://scholar.google.com/citations?user=SZgQLM4AAAAJ&hl=id&oi=ao

Azamat Serek works as Assistant Professor in Kazakh-British Technical University (KBTU) in Almaty, Kazakhstan. He has degree of PhD in Computer Science. His h-index in Scopus is 5. His research interests include artificial intelligence, deep learning, natural language processing, IT in education. He is also actively involved in curriculum development for AI and data science education. Previously he worked as senior lecturer in SDU University. He also has industrial experience in Chocofamily Holding.
Email: a.serek@kbtu.kz
Scopus profile: https://www.scopus.com/authid/detail.uri?authorId=57207763595

Bauyrzhan Berlikozha is senior lecturer in SDU university. He holds an MSc from SDU University in the Computer Science field. His research interests encompass machine learning, deep learning, IT in education.
Email: bauirzhan.berlikozha@sdu.edu.kz
Scopus profile: https://www.scopus.com/authid/detail.uri?authorId=59702974900

Babur Rustauletov is a data engineer and researcher at Suleyman Demirel University (Kazakhstan). Work focuses on data streaming, change data capture (CDC), BigQuery, and real-time analytics. Also involved in the development of data platforms and digital transformation initiatives in higher education and business.
Email: babur.rustauletov@sdu.edu.kz
Google scholar profile: https://scholar.google.com/citations?view_op=list_works&hl=ru&user=wuoIXp0AAAAJ