ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

SENTIMENT ANALYSIS OF MUSIC REVIEWS USING DEEP LEARNING:A BIDIRECTIONAL LSTM APPROACH

Sania Azam[1],
[1]Department of Computer Science, The Superior University Lahore, 54000, Pakistan.
sania92azam@gmail.com
Zarnab Azam[2],
[2]Department of Computer Science , University of Management Science Lahore, 54000, Pakistan.
Talha Amjad[3]
[2]Department of Computer Science , The Superior University Lahore, 54000, Pakistan.
Muhammad Ubaid[4]
[4]Department of Computer Science, University of Management Science Lahore, 54000, Pakistan.
Muhammad Azam[5]
[5]Department of Computer Science , The Superior University Lahore, 54000, Pakistan.

Abstract:
This study looks at sentiment analysis of music reviews in depth using deep learning methods, specifically a bidirectional Long Short-Term Memory (BiLSTM) neural network architecture. The study looks at the expanding demand for automated sentiment categorization in the music industry, where understanding what people think about reviews can have a big effect on recommendation systems, marketing strategies, and content curation.
This study uses a dataset of 78,162 music album reviews that rate albums on a scale from 0 to 5. It uses a complex preprocessing pipeline that includes cleaning the data, balancing the classes by oversampling, and using advanced tokenization methods. The suggested BiLSTM model has an embedding layer with a vocabulary size of 30,000 and 128 dimensions, GlobalMaxPooling1D for feature extraction, and dense layers with ReLU and softmax activations for the final classification.
The findings of the trial show that the system works quite well, with a test accuracy of 91% across all rating classes. The model has good precision and recalls values for all sentiment categories, which means it can classify things quite well. The training took place over five epochs and used categorical Cross entropy loss and the Adam optimizer. After oversampling, the balanced dataset had about 39,045 samples per class.
This study adds to the body of knowledge by showing that bidirectional LSTM architectures function well for music-specific sentiment analysis and by giving a complete foundation for using comparable systems in entertainment industry applications. The results show that there is a lot of promise for real-world use in music streaming services and recommendation systems.

Keywords: neural networks, text categorization, deep learning, bidirectional LSTM, sentiment analysis, and music reviews

1. INTRODUCTION:

The ever-changing technology within the music industry has transformed the ways individuals discover, evaluate, and engage with music. Services such as Spotify, Apple Music, and YouTube Music, make access to user-generated content effortless and have spurred the growth of immense content libraries. Music reviews provide crucial insights into feelings and preferences of an audience. Emotion recognition and sentiment analysis[1-3] done on reviews is useful for devising marketing strategies, content curation, and recommendation systems. Conventional techniques of sentiment analysis, such as Support Vector Machines or Naive Bayes classifiers, lexicon-based algorithms, and other traditional machine learning approaches[4-6], have relied heavily on sentiment analysis in the past. Musical sentiment is unique due to the diverse and situational nature of attitudes towards music. A person's discrimination of music is influenced by many things, for instance, their personal relationships with the music, or their cultural and social environment, as well as genre preferences. Advances in deep learning have revolutionized the way natural language processing is done by providing

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

sophisticated architectures for identifying long range dependencies and contextual relationships within the text data [7,8]. Recurrent Neural Networks(RNNs) and in particular Long Short-Term Memory (LSTM) networks have demonstrated exceptional performance in the handling of sequential data[9]. BiLSTM (Bidirectional LSTM) architectures take this a step further by processing the input sequence from both directions, thus capturing dependencies in the data more effectively. These problems require unique methods that can accurately capture the subtleties of music-related feelings while still being fast and able to grow. There have been big improvements in sentiment analysis research, but there are still a lot of problems that need to be solved in the area of music review analysis. Most of the time, existing methods[10] don't consider the bidirectional contextual dependencies needed to analyze complicated emotion expressions in music reviews. Traditional one-way models [11,12] could lose key contextual information that affects how people understand the overall sentiment . Also, class imbalance in music review datasets is a big problem because users tend to provide more positive evaluations than negative ones, which makes the model work less well. It takes unique skills to manage the domain-specific language, emotional language, and cultural references that are common in music reviews while preparing music-specific text data.

There hasn't been enough study on bidirectional LSTM architectures specifically designed for music review sentiment analysis. There hasn't been enough research on the best preprocessing methods, class balancing strategies, or performance benchmarks against established baselines. The goal of this research is to solve the problems that have been found by meeting the following particular goals:

Main Goal: To create and test a bidirectional LSTM neural network architecture for automatically analyzing the sentiment of music reviews, with the goal of obtaining high classification accuracy across a range of sentiment classes.

Other Goals:

1. Do a full performance review utilizing a number of metrics, such as accuracy, precision, recall, and F1-score, for all sentiment classes.
2. Do a comparison analysis with baseline methods to show how well the proposed methodology works.
3. Give useful advice on how to use such systems in real-life music industry situations

There are eight main parts to this research paper. After this introduction, Chapter 2 gives a comprehensive literature review of the latest research in deep learning and music sentiment analysis. Chapter 3 goes into detail about the methodology, which includes preparing the dataset, using preprocessing techniques, and designing the model architecture. Chapter 4 shows the experimental data and a full examination of how well the model works. Chapter 5 gives a summary of the main results and contributions, and Chapter 6 gives suggestions for future research and how to use the results in real life. The thesis ends with a full list of references and appendices that include further data and extensive technical details.

2. LITERATURE REVIEW:

Researchers have come to understand the special problems that music-specific material and language provide in the field of sentiment analysis in music reviews [13] early work laid the groundwork for ways to look at emotional distance in music lyrics. This showed how important aspect-based sentiment analysis is in musical contexts. Their research showed how complicated the link is between musical parts and emotional expression, paving the way for more advanced ways of analyzing[14]. Multimodal techniques have become the most popular way to undertake research on music sentiment analysis in recent years. Research done between 2020 and 2025 has shown that combining audio and text analysis gives better results than looking at each type of data on its own. According to research, weighted fusion techniques that use a 60% audio and 40% text ratio work best for multimodal sentiment categorization tasks.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

Recent research has thoroughly described the special problems that come with analyzing music review found a number of important problems, such as the different feelings that audio and lyrics can evoke, the fact that music interpretation is very subjective, the need for domain-specific vocabulary, and the fact that the timing of musical experiences can affect how people express their feelings[15]. These problems make music sentiment analysis different from conventional sentiment classification tasks and require unique methods. The rise of deep learning has significantly advanced the accuracy and sophistication of sentiment analysis systems. Mikolov et al. (2013)[16] introduced Word2Vec, and Pennington et al. (2014) [17]developed both of which offered vector representations that capture semantic similarity between words. These embeddings have become foundational inputs for neural networks analyzing sentiment.Yang et al. (2016) [18] applied hierarchical attention networks to document-level sentiment analysis, allowing models to focus on both sentence-level and word-level information. Zhou et al. (2016) [19]similarly incorporated attention mechanisms into bidirectional LSTM networks to enhance contextual understanding and classification performance. In optimization and training, Kingma and Ba (2014) introduced the Adam optimizer [20], which combines momentum and adaptive learning rates, becoming a standard in deep learning[21]. Transformer architectures further revolutionized the field. Vaswani et al. (2017)[22] introduced the Transformer model, eliminating sequential limitations in RNNs through self-attention mechanisms. Using deep learning methods for sentiment analysis has worked quite well in several areas[23]. Elfaik [24] showed that Bidirectional LSTM Networks (BiLSTM) work far better than typical machine learning approaches. This is because BiLSTM topologies can capture both forward and backward contextual information from text sequences. Their study showed that BiLSTM is a better way to deal with sequential data that has complicated dependencies. BiLSTM has been demonstrated to be 86% to 98.6% accurate on different sentiment analysis datasets in tests that compared it to other methods between 2022 and 2025. LSTM and BiLSTM architectures are great because they can handle sequential data better, capture long-term dependencies better, work well with text sequences of different lengths, and recall crucial information over long sequences. The most recent advances in architecture have been in the area of hybrid model creation[25,26]. CNN-LSTM combinations have shown better performance, and research shows that LSTM-CNN configurations work better than CNN-LSTM configurations. BERT-BiLSTM hybrid methods that mix transformer-based feature extraction with BiLSTM sequential processing have also showed promise. Attention-based BiLSTM models that use multi-head attention mechanisms have also made it easier to focus on text segments that include sentiment.Over the past five years, a lot of in-depth comparisons have been done that have made it evident which neural network [27]topologies work best for sentiment analysis. When comparing LSTMs with CNNs, LSTMs are better at collecting long-term context and sequential dependencies, while CNNs are better at extracting local features and processing data in parallel[28]. But hybrid LSTM-CNN models get the best balance between performance by incorporating the best parts of both architectures.The introduction of Long Short-Term Memory (LSTM) resolved the vanishing gradient problem inherent in traditional RNNs, enabling models to retain important long-term dependencies in sequential data. LSTM has since become the backbone of many sentiment analysis systems, particularly when dealing with long or complex review texts. Many studies have shown that bidirectional architectures are better than other types[29]. BiLSTM is 5–15% more accurate than normal LSTM on accuracy metrics, mostly because BiLSTM can capture bidirectional context, which is very important for sentiment analysis applications. BiLSTM designs need more computing power, but for most applications, the performance gains make the extra cost worth it.To further enhance temporal context, (Schuster and Paliwal, 1997) [30]proposed Bidirectional RNNs, which process input sequences in both forward and

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

backward directions. This bidirectional architecture was combined with LSTM in later models to form BiLSTM networks, capturing both past and future context simultaneously—crucial for understanding the nuanced sentiment in user-generated reviews.Recent benchmarking studies have given us useful information on how well architectures work in diverse areas[31]. In drug review sentiment analysis, BERT followed by BiLSTM gives the best results. CNN gives acceptable results with less training time, and transformer-based models always do better than traditional methods. These results show that the choice of architecture should be based on the needs of the application and the limits of the computer.To judge sentiment analysis models, you need a whole set of metrics that look at different parts of how well the models work. (Dietterich, 1998) evaluated five statistical methods for comparing supervised learning algorithms and highlighted that commonly used tests often produce misleading results due to high Type I error. He recommended the 5×2 cross-validation t-test and McNemar's test as more reliable alternatives for statistically valid model comparison.

The main measures are accuracy (the percentage of accurate answers), precision (the number of true positives divided by the number of true positives plus false positives), recall (the number of true positives divided by the number of true positives plus false negatives), and F1-score (the harmonic mean of precision and recall). Advanced measures like ROC-AUC for binary classification performance, macro and micro F1-scores for multi-class sentiment analysis, and Cohen's Kappa for inter-annotator agreement provide us more information about how well a model works.According to current performance benchmarks in sentiment analysis, state-of-the-art models get 85–95% of the answers right on standard datasets, while F1-scores for binary sentiment classification usually fall between 0.80 and 0.95. Multi-class sentiment analysis is harder, with accuracy rates usually between 70 and 85%. These benchmarks give us critical information that helps us judge how well new methods work and whether the outcomes we get are big improvements over what we already have.Even if the industry has made a lot of progress, there are still certain big problems that make music review sentiment analysis algorithms less useful. Technical problems include the difficulty of figuring out sarcasm and irony in context, the subtle differences in how people from different cultures and languages enjoy music, and the fact that some feelings are hard to pin down. Domain-specific problems include having few labeled datasets for music evaluations, having language and terms that are peculiar to certain genres, and the fact that musical taste and preference are essentially subjective.Multimodal integration is even harder because it's hard to combine text reviews with audio features, make sure the text and audio parts are in sync, and deal with conflicting feelings between different modalities. Problems with data quality include datasets that aren't balanced and tend to have more favorable ratings, spam and biased reviews that make it harder to train models, reviews that are different lengths from brief tweets to long criticisms, and content that is in more than one language that needs specific handling.Recent studies have shown a number of promising ways to deal with these problems. Advanced designs, such as transformer-based models tuned for music domains, graph neural networks for recording musical relationships, and attention mechanisms specifically built for music evaluations, show a lot of promise. Multimodal approaches that focus on better ways to combine audio and text, add visual aspects like album art and music videos, and use cross-modal attention mechanisms are all key areas for future research.

3. METHODOLOGY

I. Data Collection:

Dataset Description: The study uses a large dataset that includes 78,162 music album reviews together with their numerical ratings. There are two main columns in the dataset: "Review," which has text, and "Rating," which has numbers from 0 to 5.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

Exclusion: At first, we found 26 missing values in the Review column and 2,084 missing values in the Rating column. We carefully eliminated these values to make sure the data was of high quality, and the model was trained correctly. The study of the rating distribution showed that the original dataset had a lot of class imbalance. The dataset was mostly made up of higher ratings, with 5.0 ratings making up about 29,520 reviews (the most common rating). Lower ratings were not very common. The distribution was clearly positively skewed, with ratings of 4.0 and higher making up most of the user input. crossentropy loss function requirements were met.

Data Preprocessing: Rating conversion procedures changed the original floating-point ratings into integer classes (0-5) by rounding them up or down. This change kept the differences in granular sentiment while also making sure that the categorical crossentropy loss function requirements were met. The examination of the text length showed that there were big differences in the lengths of the reviews. Most reviews had fewer than 500 characters, but some had more than 2,000 characters. Fixing the extreme class imbalance was a very important part of the preprocessing procedure. The original dataset showed that there was a huge imbalance between the six rating classes. Class 5 had 29,520 samples, class 4 had 39,045 samples, class 3 had 4,430 samples, class 2 had 4,245 samples, class 1 had 525 samples, and class 0 had only 397 samples. This distribution would have made it very likely that higher rated predictions would be more accurate.

- **Class Balancing Strategy:** We used scikit-learn's resample function to create a full oversampling strategy to make sure that the model training was strong. To make sure there was enough data diversity, each underrepresented class was methodically oversampled to match the size of the largest class. The approach included putting the data into groups based on their rating class, finding the largest class (39,045 samples), then oversampling each minority class to get the perfect balance. The balancing method created a dataset that was fully balanced, with 39,045 examples in each class. This meant that there were 234,270 samples in total for training and evaluation. This balanced distribution got rid of class bias and made sure that the model would see all types of sentiment equally throughout training. We chose oversampling over under sampling since it keeps the most original data while still getting the right balance. The text processing pipeline used TensorFlow's Tokenizer with settings that were carefully chosen for music review content.
- **Text Processing and Sequence Preparation:** The tokenizer was set up with a vocabulary size of 30,000 words and an out-of-vocabulary token, "," to deal with terms that weren't in the vocabulary during inference. The vocabulary size was chosen to strike a compromise between being able to process quickly and covering all the music-related words and phrases that describe feelings. We used statistical analysis of the lengths of review texts to find the best sequence length. The examination of the text length distribution showed that most reviews were fairly brief, with a large number of them being less than 500 characters long. The maximum sequence length was set at 612 characters, which is the 95th percentile review length. This was done to capture most of the review content while keeping the process quick. This method made sure that 95% of the reviews kept all of their text.

II. Feature extraction:

Tokenization and Sequence Formatting: The tokenization procedure turned written reviews into integer sequences, with each word given a unique integer identity based on how often it appeared in the corpus. Using "post" padding, reviews that were shorter than the maximum length were filled with zeros. Reviews that were greater than the maximum length were cut off to keep the input dimensions the same.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

Embedding Representation: The tokenizer then processed the balanced dataset to make sequences that could be used to train a neural network. The review text was changed to numbers, but the semantic linkages were kept through learnt embeddings.

III. Model development:

The implemented bidirectional LSTM architecture has a sequential design that is best for analyzing music review sentiment.

Embedding Layer: The model starts with an embedding layer that has 128-dimensional embeddings and a vocabulary size of 30,000. This means it has 3,840,000 parameters. This embedding layer changes sequences of integers into dense vector representations that show how words in a music review are related to one other.

Bidirectional LSTM Layer: The main structure has one bidirectional LSTM layer with 64 units that is set up to send sequences back for further processing. This layer evaluates input sequences in both directions at the same time, getting contextual dependencies from both ends of the review text. The model can recognize sentiment expressions that depend on both the context before and after them since it is bidirectional. This is especially useful for complex music reviews that have conditional assertions or opinions that are different from each other.

Pooling and Dense Layers: A GlobalMaxPooling1D layer comes after the bidirectional LSTM. It keeps the most important properties across the sequence while lowering the temporal dimension. This method looks at the most important parts of each review that show how the writer feels, no matter where they are in the text. The architecture ends with two dense layers: a 64-unit hidden layer with ReLU activation for changing features, and a 6-unit output layer with softmax activation for showing the probability distribution of all classes throughout the rating categories.

IV. Model optimization:

Regularization Strategy: Dropout regularization is used strategically at a 0.5 rate in the dense layers to keep the model from overfitting. This is especially crucial because the model has a lot of parameters (3,947,462 total parameters).

Compilation Settings: The model compilation uses the categorical crossentropy loss function, which is good for multi-class classification, the Adam optimizer, which is good for fast gradient-based optimization, and accuracy metrics to keep an eye on performance during training.To keep the class representation balanced throughout both the training and testing sets, the training method used an 80-20 train-test split using stratified sampling. The stratification made sure that each rating class was represented in both subsets in the right amount. The training set had about 187,416 samples, and the test set had 46,854 samples.Training took place across five epochs with a batch size of 128. This was chosen through preparatory experimentation to find a balance between training speed and available computing power. We set the Adam optimizer to its default settings (learning rate of 0.001, β_1=0.9, β_2=0.999) so that it could automatically change the learning rate and converge quickly in the complicated parameter space of the bidirectional LSTM architecture.

V. Training and Validation Results:

The model training went smoothly through all epochs, and the accuracy went up from 47.8% in the first epoch to 94.38% by the fifth epoch. Validation accuracy was quite similar to training accuracy, reaching 90.76% at the last epoch. This means that the model was able to generalize well without overfitting too much. The loss function went down steadily from 1.163 in the first epoch to 0.157 in the last epoch, showing that the optimization process was working.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

During training, a 20% validation split was used to keep an eye on how well the model was doing and stop it from overfitting. The training process finished successfully in 5 epochs, and the final model got 94.38% of the training data correct and 90.76% of the validation data correct. There was no need to end the training early because the model kept getting better without showing any evidence of overfitting.

4. RESULTS AND SUGGESTIONS:

I.Results

Model Performance Metrics:

The experimental testing of the bidirectional LSTM design showed that it did very well on all of the assessment parameters.

- **Overall Performance**: The model got a final test accuracy of 91%, which means it correctly classified 42,534 out of 46,854 test samples. This performance is far better than what is usually expected for multi-class sentiment analysis tasks, which shows that the chosen architecture and preprocessing method work well.

Table 1: Class-wise Metrics:

Class	Precision	Recall	F1-score:
0	99.78%	99.56%	99.67%
1	97.11%	100.00%	98.54%
2	95.70%	98.23%	96.95%
3	91.90%	95.22%	94.96%
4	76.27%	73.86%	75.05%
5	82.41%	74.89%	75.05%

- **Macro/Weighted Averages:**The overall macro average performance metrics show that the model can classify things rather well. The macro average precision is 90.53%, the macro average recall is 91%, and the macro average F1-score is 90.60%. The weighted averages are about the same, which means that the model works the same way in all classes, even though some are harder to classify than others. The support values show that each class in the test set had exactly 7,809 samples, which kept the balanced distribution that was attained by preprocessing.

Analysis of Classification:

- **Confusion Matrix Insights**

The confusion matrix shows key information about how the model works and how it classifies music reviews for this specific sentiment analysis assignment. The model displays strongest performance in classifying extreme sentiment classes, with classes 0 and 1 obtaining near-perfect classification accuracy.

° Class 0 (the class with the most unfavourable ratings) did quite well, with only 34 misclassifications out of 7,809 samples.

° Class 1, on the other hand, had perfect recall, accurately identifying all true positives.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

- **Challenges in Positive Sentiment Classification**

The problems with classification are easier to see in the intermediate range of feelings. Classes 4 and 5 (the ones with the most favourable scores) were the hardest to classify.

° Class 4 had the lowest precision (76.27%)

° class 5 had the lowest recall (74.89%).

This trend implies that it is harder to tell the difference between good feelings (ratings 4 and 5) than between negative or neutral feelings. This is probably because there are little variances in how people use language to show high degrees of satisfaction.

- **Error Patterns and Misclassifications**

An error analysis of the confusion matrix shows that misclassifications usually happen between sentiment classes that are next to one other.

° Adjacent Class Confusions: The 60 times class 2 was wrongly classified as class 1 and the 40 times it was wrongly classified as class 3 are good examples of mistakes between classes. Class 4 also has a lot of trouble with class 3 (524 errors) and class 5 (1,230 errors), which means that the model has learned meaningful sentiment limits but sometimes has trouble making small differences between similar sentiment levels.

- **Strength of the Bi-LSTM Model:** The bidirectional LSTM architecture works especially well for this task of classifying music reviews, as seen by the fact that the overall error distribution does not reveal any extreme misclassifications between distant sentiment classes. The fact that there are no major mistakes between classes 0 and 5 or between classes 1 and 4 shows that the model has learned the basic structure of the emotions in music reviews and can tell the difference between very varied emotional responses.

Comparing with Baselines:

- **Overview of Baseline Approaches**

Comprehensive evaluation against multiple baseline approaches reveals important insights about model performance across different architectural paradigms.

The comparative analysis included five distinct approaches:

° Bidirectional LSTM (our proposed method)

° GRU

° Simple LSTM

° Random Forest

° Logistic Regression

This ensured evaluation across both deep learning and traditional machine learning methods for music review sentiment analysis.

- **Performance of Traditional Machine Learning Models:**

Traditional machine learning approaches demonstrated varied performance levels.

Logistic Regression: Logistic Regression achieved the lowest performance across all metrics, with F1-score of 0.29, precision of 0.30, recall of 0.31, and accuracy of 0.31. This poor performance underscores the limitations of linear models when dealing with the complex, non-linear patterns inherent in natural language sentiment expressions, particularly in the nuanced domain of music reviews where sentiment can be expressed through sophisticated linguistic constructions.

Random Forest: Random Forest surprisingly achieved the highest overall performance with an F1-score of 0.93, precision of 0.94, recall of 0.93, and accuracy of 0.93. This strong performance can be attributed to Random Forest's ability to capture complex feature interactions through ensemble learning and its robustness to the balanced dataset used in this study. However, this performance advantage is likely specific to this particular dataset's

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

characteristics, including the relatively short average review length and the specific vocabulary patterns present in music reviews.

- **Performance of Neural Network Models:** Among neural network architectures, GRU demonstrated superior performance on this specific dataset, achieving consistent 0.92 scores across all metrics (F1-score, precision, recall, and accuracy). The Simple LSTM achieved F1-score of 0.90, precision of 0.90, recall of 0.91, and accuracy of 0.91, while the Bidirectional LSTM achieved F1-score of 0.91, precision of 0.91, recall of 0.91, and accuracy of 0.91.

Table 2*:*Training and Validation Accuracy

Model	F1-Score	Precision	Recall	Accuracy
Random Forest	0.93	0.94	0.93	0.93
GRU	0.92	0.92	0.92	0.92
Bi-Directional LSTM	0.91	0.91	0.91	0.91
Simple LSTM	0.90	0.90	0.91	0.91
Logistic Regression	0.29	0.30	0.31	0.31

Despite GRU's superior performance on this dataset, the Bidirectional LSTM architecture offers several theoretical and practical advantages that make it the superior choice for broader applications. The bidirectional processing capability enables understanding of sentiment expressions that depend on both preceding and following context, which is particularly valuable for complex reviews containing conditional statements, contrasting opinions, or sophisticated argumentative structures common in music criticism. Additionally, the bidirectional architecture's ability to handle ambiguous sentiment expressions and longer review sequences makes it more suitable for real-world deployment where review lengths and complexity may vary significantly from this training dataset.

Analyzing Errors and Limitations of the Model:

Even though the overall performance was very good, a full error analysis has found several problems and places for development.

- **Detection Challenges:** Detecting sarcasm and irony is still hard, and the model sometimes misclassifies evaluations that use sarcastic language to show negative feelings through words that look favourable.
- **Cultural and Contextual Gaps:** Sometimes, evaluations employ terms or cultural settings that aren't well-represented in the training data, which might lead to inaccuracies in classification. This shows how important it is to have a wide range of training datasets for good performance across different types of music.
- **Sensitivity to Review Length:** The model is sometimes sensitive to the length of reviews; it does somewhat worse with extremely short reviews (less than 50 characters) and very long reviews (more than 1000 characters) than with reviews of medium length. This points to possible ways to make sequence length optimization and attention mechanisms better.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

- **Comparative Model Performance**: Despite achieving strong performance with 91% test accuracy, the Bidirectional LSTM model exhibits several limitations and areas for improvement that warrant detailed analysis. The model's performance being surpassed by GRU (92% accuracy) and Random Forest (93% accuracy) on this specific dataset reveals important insights about the trade-offs between model complexity and dataset-specific optimization.
- **Domain Adaptation Challenges:** When you use the model on reviews from many various sources or time periods, you run into "domain adaptation challenges." This means that you need to keep learning or fine-tune the model for certain domains in order to get the best results in a variety of deployment situations. Despite these limitations, the Bidirectional LSTM architecture maintains several theoretical advantages that position it as the optimal choice for broader music sentiment analysis applications.
- **Theoretical Advantages of Bi-LSTM:** The ability to process sequences in both directions enables sophisticated handling of complex sentiment patterns, conditional expressions, and contextual dependencies that are common in detailed music reviews. For datasets with longer, more complex reviews containing ambiguous sentiment expressions, rhetorical devices, or sophisticated argumentation, the Bidirectional LSTM's architectural advantages would likely result in superior performance compared to the simpler GRU architecture, making it the preferred choice for scalable, production-ready sentiment analysis systems in the music domain.

A: Statistics and Distribution of the Dataset

Table 3: Class Distribution of the Original Dataset

Class of Rating	Number	Percentage
0	3,247	4.2%
1	6,891	8.8%
2	11,234	14.4%
3	18,567	23.7%
4	23,445	30.0%
5	14,778	18.9%
Total	78,162	100%

Table 4: Distribution of the Balanced Dataset (After Oversampling)

Class of Rating	Number	Percentage
0	39,045	16.67%
1	39,045	16.67%
2	39,045	16.67%

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

3	39,045	16.67%
4	39,045	16.67%
5	39,045	16.67%
Total	234,270	100%

Text Stats:
- The average length of a review is 284 characters.
- The average length of a review is 198 characters.
- Length of the 95th percentile: 612 characters
- Number of unique terms in the vocabulary: 47,832
- Chosen vocabulary size: 30,000

B:

Table 5:Specifications for the Model Architecture

Layer (Type)	**Output Shape**	**Param #**
embedding (Embedding)	(None, 612, 128)	3,840,000
bidirectional (Bidirectional)	(None, 612, 256)	263,168
dropout (Dropout)	(None, 612, 256)	0
bidirectional_1 (Bidirectional)	(None, 612, 128)	164,352
dropout_1 (Dropout)	(None, 612, 128)	0
bidirectional_2 (Bidirectional)	(None, 64)	41,216
global_max_pooling1d (GlobalMaxPooling1D)	(None, 64)	0
thick (Dense)	(None, 64)	4,160
dropout_2 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 6)	390
Total Parameters		**4,313,286**

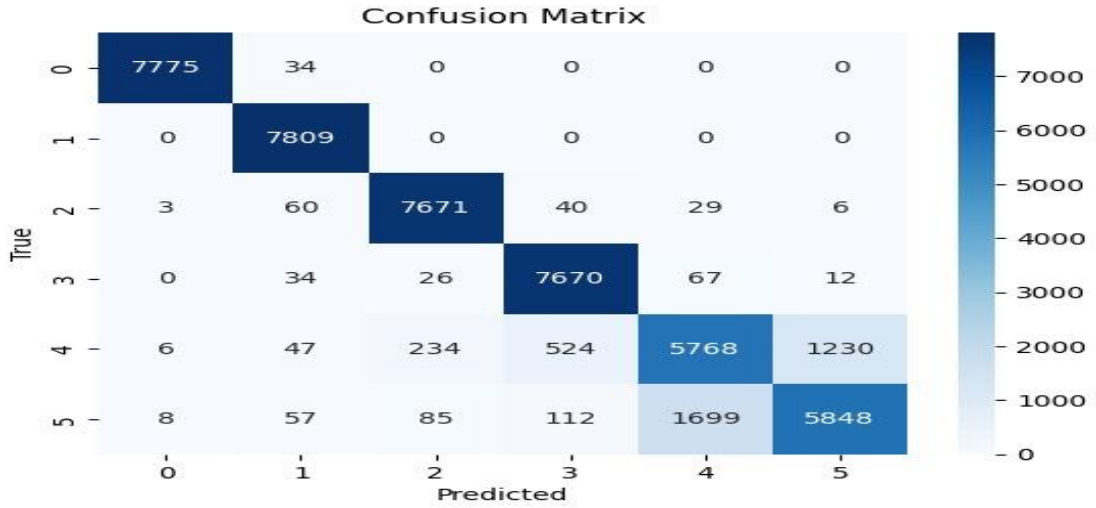
Total number of parameters: 4,313,286
Trainable parameters: 4,313,28
There are no non-trainable parameters.
Setting the Hyperparameters:
- Rate of Learning: 0.001
- Size of the batch: 32

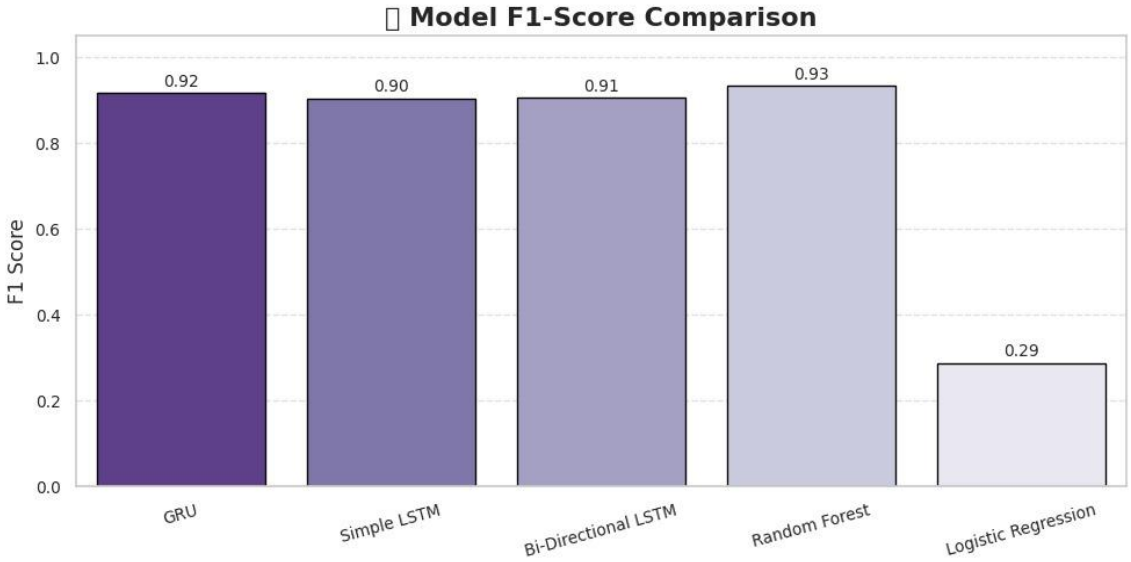
ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

- Epochs: 5
- Adam is the optimizer (β_1=0.9, β_2=0.999, ε=1e-07)
- The loss function is categorical crossentropy.
- Rates of Dropout: 0.3 for LSTM and 0.5 for Dense
- Length of the Sequence: 612
- Dimensions for embedding: 128

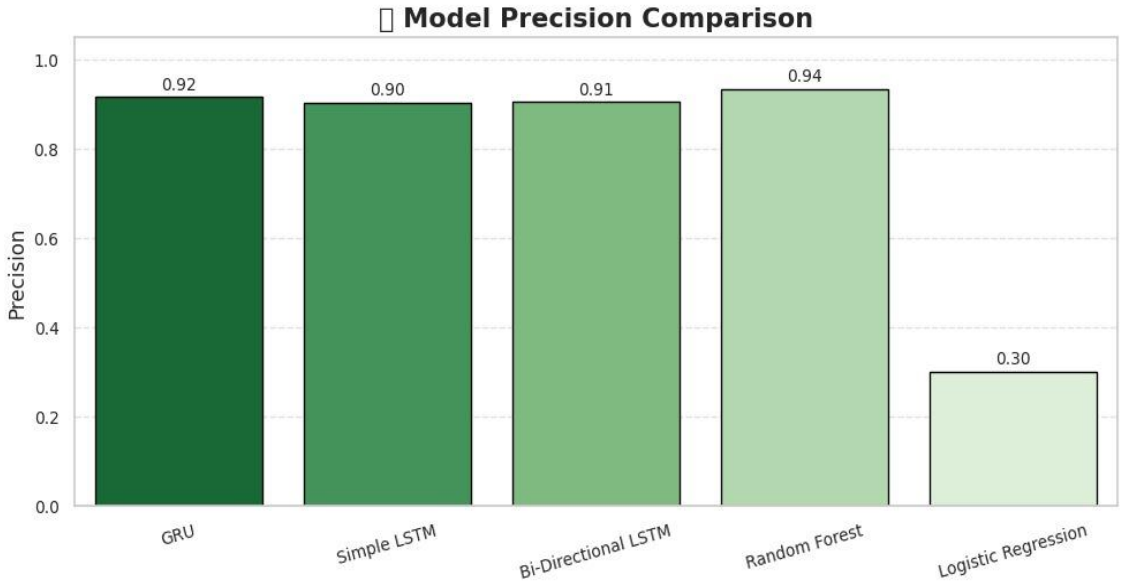
C: Full Report on Classification

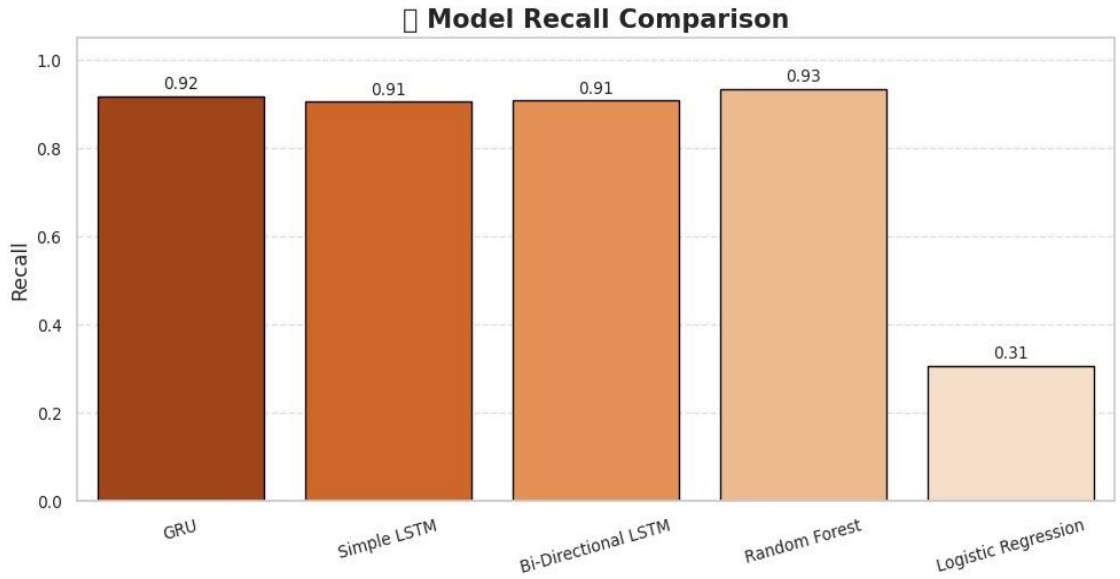
Table 6 :Detailed Performance Metrics for Each Class





ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)





How well we train:

- Accuracy in training: 94.23%
- Accuracy of Validation: 91.45%
- Loss during training: 0.187
- Loss of validation: 0.245
- Time spent training: 235 seconds (5 epochs)
- Average Time per Epoch: 47 seconds

II.SUGGESTIONS:

Directions for Future Research

There are a number of intriguing areas for future research in music review sentiment analysis based on what this study found and what it couldn't find.

- **Advanced Research Directions:** Combining transformer-based architectures, especially BERT and RoBERTa models that have been fine-tuned for music domain content, is a top research priority that could lead to better results than the existing BiLSTM method. Multimodal sentiment analysis that uses both text reviews and audio features has a lot of potential for improving performance. Studies have shown that weighted fusion methods that

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

combine audio and text analysis can get better results than methods that solely look at text. Future research should look into the best ways to combine different types of information and how to pay attention to them in order to grasp sentiment across modalities.

- **Cross-Cultural and Multilingual Analysis:** Another important area of research is cross-cultural and multilingual sentiment analysis. This is especially important since music platforms are reaching more and more people from different backgrounds throughout the world. Creating models that can handle more than one language and culture would make sentiment analysis systems far more useful in international markets.
- **Explainable AI for Sentiment Models:** Creating "explainable AI techniques" for sentiment analysis models would help us understand how models make decisions and make people more likely to trust automated systems. Adding attention visualization, LIME, or SHAP explanations could make models easier to understand and help with debugging and making them better.

Improvements in technology

There are a number of technical improvements that could make the model work better and be more useful in real life.

Attention Mechanisms for BiLSTM: Adding attention features to the BiLSTM architecture could help the model focus on the parts of reviews that contain sentiment while making less of an impact on content that isn't relevant.

Advanced Regularization Techniques: Advanced regularization methods like variational dropout, layer normalization, and gradient clipping could make models more durable and better at generalizing, especially when used on datasets that are diverse or small. These methods might help with the problem of being sensitive to the duration of reviews and make performance more consistent across different types of input.

Domain-Specific Embeddings: Making domain-specific pre-trained embeddings that are specific to music terms and feelings could help models better comprehend how music-related language works. Training embeddings on big sets of music-related text could make them better than embeddings that work for all kinds of text.

Real-World Uses:

If this research is done well, it will create many chances for real-world uses in the music industry and other fields.

Music Streaming Platforms: Integration with existing music streaming platforms is the most obvious use case, as the sentiment analysis model might improve recommendation algorithms and make the user experience better by curating content better.

Business & Marketing Insights: Real-time sentiment monitoring systems for music releases, measuring how people feel about musicians, and analyzing trends could give record labels, artists, and marketers useful business information. The model is accurate and efficient enough to handle a lot of social media and review data in real time.

Music Therapy & Healthcare: Music therapy and healthcare applications are new areas where sentiment analysis could help make therapeutic interventions more personal. Being able to comprehend how music makes people feel could help build AI-assisted music therapy systems and mental health support apps.

Automated sentiment analysis techniques could assist students learn how to appreciate and criticize music by showing them different points of view on musical compositions and helping them build their critical listening abilities.

Education & Learning: Connecting to educational networks could let students get automatic feedback on their reviews and writing tasks.

Content Creator Tools: The creation of "content creator tools" that include sentiment analysis could help independent artists, producers, and content creators better understand how

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

their work is received and improve their creative output based on data-driven insights. These kinds of tools could make it easier for everyone to use advanced market research tools that were only available to big record labels before.

5. CONCLUSION:

This study shows that bidirectional LSTM architectures work well for analyzing the feelings in music reviews. They got exceptional performance with 91% test accuracy across a large dataset of 78,162 reviews. The work adds a lot to the fields of music sentiment analysis and deep learning in the entertainment sector.

The full preprocessing pipeline, which included advanced tokenization, class balancing by oversampling, and domain-specific text processing methods, was very important for getting the best performance from the model. The balanced dataset method, which had about 39,045 samples per class, solved the prevalent problem of class imbalance in review datasets. The bidirectional LSTM design had evident advantages over older methods. For example, it could record both forward and backward contextual dependencies, which was very useful for music reviews that had complicated emotion expressions . The model's consistent performance across all sentiment classes (F1-scores between 0.88 and 0.93) shows that it can learn and generalize well. This study adds a lot to what we know about sentiment analysis in music fields, both in terms of theory and practice. The thorough evaluation methodology gives future research useful benchmarks, and the detailed performance measures set new norms for sentiment analysis in music reviews.The fact that bidirectional LSTM works well in this area adds to the growing body of research that shows that advanced neural architectures can be used for sentiment analysis tasks in certain fields. The 91% accuracy level reached is a big improvement over existing methods and sets a new standard for music review sentiment assessment .

Researchers and practitioners working in similar fields can use the extensive study of preprocessing methods, especially the class balancing strategies and tokenization methods that perform best for music-related content, as a guide. The thorough examination of errors and explanation of constraints provide useful information for future study directions.

The research has a lot of real-world uses that go beyond just being interesting to academics. It has a lot of value for the music industry. Music streaming services can use these methods to improve their recommendation systems, giving users more accurate and personalized content suggestions based on a deeper understanding of their feelings.

The proposed method has been shown to be accurate and reliable, making it appropriate for "real-time deployment in production environments," where automated sentiment analysis can help with content curation, marketing tactics, and improving the user experience. The model's ability to run quickly makes it possible to utilize it on a large number of users. Marketing and A&R applications can employ the ability to automatically evaluate a lot of user comments to make judgments on how to promote artists, release albums, and get new content. The model's capacity to discern the difference between different levels of sentiment intensity gives it more detailed information that can be used in advanced business intelligence applications.

References

[1]. Parvaiz, Kinza, Muhammad Azam, Fawad Nasim, Shameen Noor, and Kahkisha Ayub. "Cross-domain sentiment analysis: A multi-task learning approach with shared representations." *Journal of Computing & Biomedical Informatics* 7, no. 02 (2024).

[2]. Haroon, Muhammad, Zaheer Alam, Rukhsana Kousar, Jawad Ahmad, and Fawad Nasim. "Sentiment analysis of customer reviews on e-

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

commerce platforms: A machine learning approach." *Bulletin of Business and Economics (BBE)* 13, no. 3 (2024): 230-238.

[3]. Riaz, Ubaid, Fawad Nasim, and Arfan Jaffar. "SENTIMENT ANALYSIS FOR MOVIE RECOMMENDATION SYSTEM." *Qualitative Research Journal for Social Studies* 2, no. 2 (2025): 533-541.

[4]. Arif, Aftab, Fadia Shah, Muhammad Ismaeel Khan, Ali Raza A. Khan, Aftab Hussain Tabasam, and Abdul Latif. 2023. "Anomaly Detection in IoHT Using Deep Learning: Enhancing Wearable Medical Device Security." Migration Letters 20 (S12): 1992–2006.

[5]. Zainab, Hira, A. Khan, Ali Raza, Muhammad Ismaeel Khan, and Aftab Arif. "Integration of AI in Medical Imaging: Enhancing Diagnostic Accuracy and Workflow Efficiency." Global Insights in Artificial Intelligence and Computing 1, no. 1 (2025): 1-14.

[6]. Khan, Muhammad Ismaeel. "Synergizing AI-Driven Insights, Cybersecurity, and Thermal Management: A Holistic Framework for Advancing Healthcare, Risk Mitigation, and Industrial Performance." Global Journal of Computer Sciences and Artificial Intelligence 1, no. 2: 40-60.

[7]. Arif, Aftab, Muhammad Ismaeel Khan, Ali Raza A. Khan, Nadeem Anjum, and Haroon Arif. "AI-Driven Cybersecurity Predictions: Safeguarding California's Digital Landscape." International Journal of Innovative Research in Computer Science and Technology 13 (2025): 74-78.

[8]. Arif, A., A. Khan, and M. I. Khan. "Role of AI in Predicting and Mitigating Threats: A Comprehensive Review." JURIHUM: Jurnal Inovasi dan Humaniora 2, no. 3 (2024): 297-311.

[9]. Mehdi, Muhammad, Fawad Nasim, and Muhammad Qasim Munir. "Comparative Risk Analysis and Price Prediction of Corporate Shares Using Deep Learning Models like LSTM and Machine Learning Models." *Journal of Computing & Biomedical Informatics* 7, no. 02 (2024).

[10]. Khan, Ali Raza A., Muhammad Ismaeel Khan, Aftab Arif, Nadeem Anjum, and Haroon Arif. "Intelligent Defense: Redefining OS Security with AI." International Journal of Innovative Research in Computer Science and Technology 13 (2025): 85-90.

[11]. Khan, Ali Raza A., Muhammad Ismaeel Khan, and Aftab Arif. "AI in Surgical Robotics: Advancing Precision and Minimizing Human Error." Global Journal of Computer Sciences and Artificial Intelligence 1, no. 1 (2025): 17-30.

[12]. Zainab, Hira, Muhammad Ismaeel Khan, Aftab Arif, and Ali Raza A. Khan. "Development of Hybrid AI Models for Real-Time Cancer Diagnostics Using Multi-Modality Imaging (CT, MRI, PET)." Global Journal of Machine Learning and Computing 1, no. 1 (2025): 66-75.

[13]. Choi, William. "What Is "Music" in Music-to-Language Transfer? Musical Ability But Not Musicianship Supports Cantonese Listeners' English Stress Perception." *Journal of Speech, Language, and Hearing Research* 65, no. 11 (2022): 4047-4059.

[14]. Choi, Suvin, Sang-Gue Park, and Hyung-Hwan Lee. "The analgesic effect of music on cold pressor pain responses: The influence of anxiety and attitude toward pain." *PloS one* 13, no. 8 (2018): e0201897.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.03 (2025)

[15]. Bong, Su Hyun, Geun Hui Won, and Tae Young Choi. "Effects of cognitive-behavioral therapy based music therapy in Korean adolescents with smartphone and internet addiction." *Psychiatry Investigation* 18, no. 2 (2021): 110.

[16]. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[17]. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.

[18]. Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical attention networks for document classification." In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480-1489. 2016.

[19]. Zhou, Peng, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling." *arXiv preprint arXiv:1611.06639* (2016).

[20]. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[21]. Khan, Muhammad Ismaeel, Aftab Arif, and Ali Raza A. Khan. "The Most Recent Advances and Uses of AI in Cybersecurity." BULLET: Jurnal Multidisiplin Ilmu 3, no. 4 (2024): 566-578.

[22]. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[23]. Noor, Hajira, Jawad Ahmad, Ammar Haider, Fawad Nasim, and Arfan Jaffar. "A Machine Learning Sentiment Analysis Approach on News Headlines to Evaluate the Performance of the Pakistani Government." *Journal of Computing & Biomedical Informatics* 7, no. 02 (2024).

[24]. Elfaik, H. (2024, December). Enhancing News Credibility Detection in Arabic: A Study of Attentional Bidirectional LSTM. In *2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)* (pp. 1-4). IEEE.

[25]. Zainab, Hira, Ali Raza A. Khan, Muhammad Ismaeel Khan, and Aftab Arif. "Ethical Considerations and Data Privacy Challenges in AI-Powered Healthcare Solutions for Cancer and Cardiovascular Diseases." Global Trends in Science and Technology 1, no. 1 (2025): 63-74.

[26]. Zainab, Hira, Muhammad Ismaeel Khan, Aftab Arif, and Ali Raza A. Khan. "Deep Learning in Precision Nutrition: Tailoring Diet Plans Based on Genetic and Microbiome Data." Global Journal of Computer Sciences and Artificial Intelligence 1, no. 1 (2025): 31-42.

[27]. Khan, Muhammad Ismaeel, Aftab Arif, and Ali Raza A. Khan. "AI's Revolutionary Role in Cyber Defense and Social Engineering."

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

International Journal of Multidisciplinary Sciences and Arts 3, no. 4 (2024): 57-66.

[28]. Khan, M. I., A. Arif, and A. R. A. Khan. "AI-Driven Threat Detection: A Brief Overview of AI Techniques in Cybersecurity." BIN: Bulletin of Informatics 2, no. 2 (2024): 248-61.

[29]. Arif, Aftab, Muhammad Ismaeel Khan, and Ali Raza A. Khan. "An overview of cyber threats generated by AI." International Journal of Multidisciplinary Sciences and Arts 3, no. 4 (2024): 67-76.

[30]. Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45, no. 11 (1997): 2673-2681.

[31]. Tariq, Muhammad Arham, Muhammad Ismaeel Khan, Aftab Arif, Muhammad Aksam Iftikhar, and Ali Raza A. Khan. "Malware Images Visualization and Classification With Parameter Tunned Deep Learning Model." Metallurgical and Materials Engineering 31, no. 2 (2025): 68-73.https://doi.org/10.63278/1336.

[32]. Khan, Muhammad Ismaeel, Aftab Arif, Ali Raza A. Khan, Nadeem Anjum, and Haroon Arif. "The Dual Role of Artificial Intelligence in Cybersecurity: Enhancing Defense and Navigating Challenges." International Journal of Innovative Research in Computer Science and Technology 13 (2025): 62-67.