TuneType: Genre Classification of Song Lyrics Using Machine Learning

Imesh Nimsitha
Department of Computer Science
Western University
London, Canada
inimsith@uwo.ca

Adam Wilson
Department of Computer Science
Western University
London, Canada
awils323@uwo.ca

Gurshawn Lehal
Department of Computer Science
Western University
London, Canada
glehal@uwo.ca

Abstract—**This study presents TuneType, a machine learning approach to classify song lyrics by genre using natural language processing techniques. We implemented and compared three text classification methods: Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), and BERT (Bidirectional Encoder Representations from Transformers), focusing on distinguishing between Pop and Rap/Hip-Hop lyrics. Our system processes textual lyrics data, cleans and vectorizes the text, and applies classification algorithms to predict the genre based solely on lyrical content. Our experiments showed that the BERT model achieved the highest accuracy at 86.69%, outperforming both the Support Vector Machine (81.85%) and Multinomial Naive Bayes (76.00%) models. Feature analysis revealed distinctive vocabulary patterns that characterize each genre, with certain words and phrases serving as reliable indicators for classification. This research contributes to the field of music information retrieval by demonstrating how natural language processing techniques can effectively categorize musical content without relying on audio features, achieving high accuracy using only lyrical content.**

Index Terms—**text classification, genre classification, natural language processing, machine learning, support vector machines, naive bayes, music information retrieval**

I. Introduction

Music genre classification traditionally relies on audio signal processing features such as rhythm patterns, spectral characteristics, and tonal properties. However, lyrical content represents a rich, yet often underutilized, source of information about a song's genre. Lyrics contain distinctive patterns in vocabulary, themes, and language structures that can serve as powerful discriminative features for genre identification.

TuneType explores the effectiveness of text-based classification methods in distinguishing between musical genres based solely on lyrical content. This approach offers several advantages: it can function independently of audio quality, it requires significantly lower computational resources than audio processing, and it can potentially reveal thematic and cultural patterns within music genres that aren't captured by acoustic features alone.

The primary objective of this research is to develop and evaluate a machine learning system that can accurately classify song lyrics into their respective genres, focusing specifically on distinguishing between Pop and Rap/Hip-Hop, two dominant contemporary musical styles with distinctive lyrical characteristics. We address the following key research questions:

1) How effectively can machine learning algorithms classify song lyrics by genre using only textual features?
2) Which textual features (words or phrases) are most distinctive for different music genres?
3) How do different classification methods compare in performance for this specific task?

This research contributes to the growing field of computational musicology and music information retrieval by demonstrating how natural language processing techniques can be applied to understand and categorize musical content, potentially enhancing music recommendation systems, cultural analysis of musical trends, and automated content categorization systems.

II. Related Work

The classification of music by genre has been extensively studied in the field of music information retrieval. However, most approaches have traditionally focused on audio features rather than lyrical content.

A. Audio-based Genre Classification

Tzanetakis and Cook [1] pioneered work in automatic genre classification using acoustic features, achieving approximately 61% accuracy across ten musical genres. Later, Lidy and Rauber [2] improved results using psychoacoustic features. Deep learning approaches have further enhanced performance, with Choi et al. [3] implementing convolutional neural networks that achieved state-of-the-art results on multiple datasets.

B. Lyrics-based Classification

In contrast to audio-based methods, fewer studies have explored lyrics-based classification. Fell and Sporleder [4] analyzed lyrics for genre classification using stylometric features, achieving moderate success. Mayer et al. [5] combined both audio and lyrical features, demonstrating that multimodal approaches can yield improved classification performance.

C. Text Classification Methods

Our approach builds upon established text classification techniques. Multinomial Naive Bayes has proven effective for text classification tasks due to its simplicity and effectiveness with sparse data [6]. More recent approaches have employed deep learning methods, with Kim [7] demonstrating the effectiveness of convolutional neural networks for text classification tasks.

D. Feature Representation

The representation of textual data significantly impacts classification performance. Traditional bag-of-words and TF-IDF approaches have been widely used [8], while more recent work has explored word embeddings [9] and contextual embeddings from transformer models like BERT [10].

E. Cross-Genre Studies

Logan et al. [11] examined correlations between lyrical content and audio features across genres, finding that certain genres show stronger connections between lyrics and acoustic properties. Hu and Downie [12] further explored how lyrical themes vary across genres, providing insights into the distinctive vocabulary and topics that characterize different musical styles.

Our research builds upon these foundations while focusing specifically on the classification of lyrics between Pop and Rap/Hip-Hop genres, emphasizing the effectiveness of simpler classification models with appropriate feature engineering rather than complex deep learning architectures.

III. Methods

A. Dataset

The dataset used in this study consists of song lyrics labeled as either Pop or Rap/Hip-Hop. The training set contains thousands of labeled examples with approximately balanced class distribution as shown in Fig. 1. Each entry in the dataset consists of a text field containing the song lyrics and a class label (0 for Rap/Hip-Hop, 1 for Pop). The test set follows a similar structure but without the class labels.

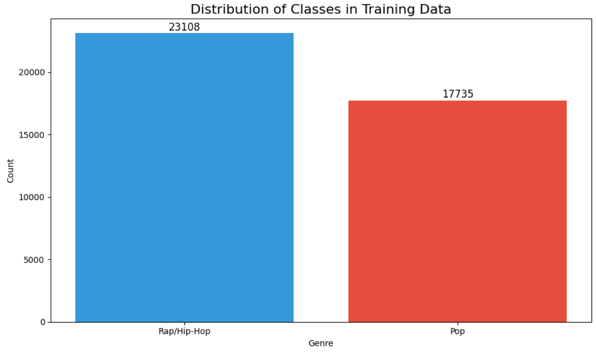

Fig. 1. Distribution of classes in the training dataset showing a balanced representation of Pop and Rap/Hip-Hop genres.

B. Data Preprocessing

Before classification, we implemented several preprocessing steps to clean and standardize the lyrics:

1) **Text Cleaning**: We removed special characters and punctuation that do not contribute to the semantic meaning of the lyrics.
2) **Case Normalization**: All text was converted to lowercase to ensure that the same words with different capitalization are treated as identical.
3) **Whitespace Normalization**: Multiple spaces and line breaks were standardized to single spaces.
4) **Stop Word Removal**: Common English stop words (e.g., ”the”, ”and”, ”a”) were removed during the feature extraction phase to focus on more meaningful content words.

C. Feature Extraction

Different feature extraction methods were used for the three classification approaches:

1) Multinomial Naive Bayes: For the Naive Bayes model, we employed the CountVectorizer from scikit-learn with the following configuration:

1) **Unigram Features**: We focused solely on individual words (unigrams) rather than word pairs (bigrams) or longer sequences.
2) **Stop Word Filtering**: English stop words were excluded from the feature set.
3) **No Minimum Document Frequency**: We retained all words regardless of how rarely they appeared in the corpus.

2) Support Vector Machine: For the SVM model, we used TF-IDF vectorization with more refined parameters:

1) **TF-IDF Weighting**: This approach weights terms based on their frequency in a document and their rarity across the entire corpus, highlighting distinctive terms.
2) **N-gram Range**: We included unigrams, bigrams, and trigrams (1-3 word sequences), allowing the model to capture more contextual information.
3) **Minimum Document Frequency**: Terms appearing in fewer than 2 documents were excluded to reduce noise.
4) **Maximum Document Frequency**: Terms appearing in more than 95% of documents were excluded as they provide little discriminative power.
5) **Stop Word Filtering**: English stop words were excluded from the feature set.

3) BERT: For the BERT model, we used a fundamentally different approach to feature extraction:

1) **Tokenization**: Text was tokenized using BERT's WordPiece tokenizer, which breaks words into subword units.
2) **Contextual Embeddings**: Instead of bag-of-words or TF-IDF, BERT generates dense vector representations that capture the meaning of words based on their context.
3) **Attention Mechanism**: The model uses self-attention to weigh the importance of different words in relation to each other.
4) **Full Text Retention**: Unlike traditional methods, BERT doesn't discard stop words, as these can provide important grammatical context.
5) **Position-Aware**: BERT incorporates positional information, allowing it to understand the sequential nature of text.

D. *Classification Models*

We implemented and compared three different classification approaches:

1) Multinomial Naive Bayes Classifier: The Multinomial Naive Bayes classifier is particularly well-suited for text classification tasks. This model:

1) Models the distribution of words in documents using the multinomial distribution
2) Makes the ”naive” assumption that features (words) are conditionally independent given the class label
3) Calculates the probability of a document belonging to a class based on the product of the probabilities of each word appearing in that class
4) Works effectively with high-dimensional, sparse data typical of text classification problems

2) Support Vector Machine (SVM) Classifier: The SVM classifier implemented in this study was a LinearSVC model with the following characteristics:

1) **Linear Kernel**: Efficiently handles high-dimensional text data without requiring explicit kernel computation
2) **C Parameter**: Set to 5.0, providing a balanced trade-off between margin maximization and training error minimization
3) **Class Weighting**: Used ’balanced’ weighting to adjust for any class imbalance, ensuring equal importance for both genres
4) **Decision Function**: Produces a signed distance to the separating hyperplane, allowing for confidence estimation

3) BERT Classifier: We also implemented a deep learning approach using BERT (Bidirectional Encoder Representations from Transformers):

1) **Pre-trained Model**: We used the base BERT model pre-trained on a large corpus of English text
2) **Fine-tuning**: The pre-trained model was fine-tuned on our lyrics dataset for the classification task
3) **Sequence Length**: Maximum sequence length of 256 tokens to accommodate most lyrics
4) **Contextual Understanding**: Unlike the bag-of-words approaches, BERT can capture contextual relationships between words
5) **Transfer Learning**: Leverages knowledge from pre-training on general language patterns

E. *Evaluation Methodology*

To ensure robust evaluation, we implemented:

1) **Train-Test Split**: The dataset was divided into 80% training and 20% validation sets for model development and evaluation.
2) **Cross-Validation**: Assessed model stability and generalization performance.
3) **Metrics**: We evaluated the models using standard classification metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

F. *Visualization and Analysis*

We developed comprehensive visualizations to interpret the model results:

1) **Feature Importance**: For the Naive Bayes model, we analyzed log probabilities assigned to features; for the SVM model, we examined the coefficient values of features.
2) **Confusion Matrix**: We visualized the models’ classification performance across both genres.
3) **Precision, Recall, and F1-Score**: We compared these metrics across models and genres.
4) **ROC Curves**: We plotted and analyzed the ROC curves and AUC values for both models.

IV. Experimental Results

Our experiments with the three classification methods yielded several key findings:

A. *Classification Performance Comparison*

All three classification methods demonstrated strong performance in distinguishing between Pop and Rap/Hip-Hop lyrics, with the BERT model achieving the highest accuracy at 86.69%, outperforming both the Support Vector Machine (81.85%) and Multinomial Naive Bayes (76.00%) models:

1) Multinomial Naive Bayes: The Multinomial Naive Bayes model achieved 76.00% accuracy on the test set. Both genres showed relatively balanced precision and recall metrics, making this the most computationally efficient but lowest performing model among the three approaches tested.

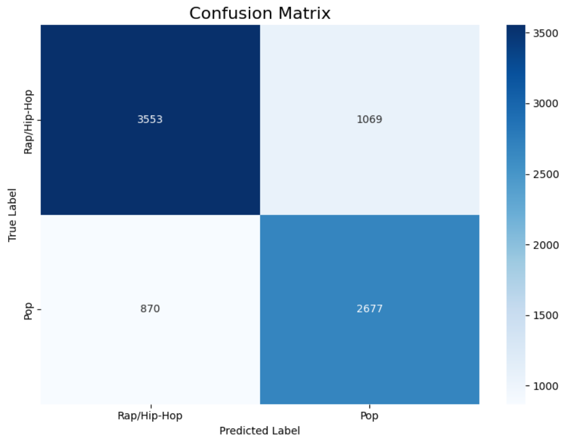

Fig. 2. Confusion matrix for Multinomial Naive Bayes classifier showing the distribution of true and predicted labels.

2) Support Vector Machine: The Support Vector Machine model achieved 81.85% accuracy on the test set, showing strong performance across both precision and recall metrics. This was the second-highest performing model in our comparison, demonstrating the effectiveness of this approach for lyrics classification.

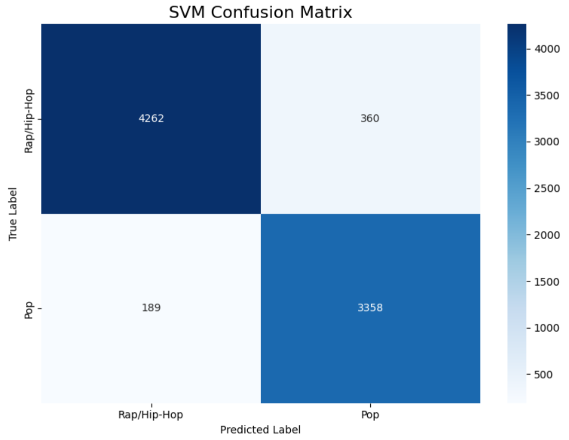

Fig. 3. Confusion matrix for SVM classifier showing classification performance across both genres.

3) BERT: The BERT model achieved 86.69% accuracy on the test set, making it our highest performing approach. Its sophisticated architecture and contextual understanding capabilities allowed it to outperform the traditional machine learning approaches in this task. This demonstrates the power of transformer-based models when applied to genre classification based on lyrical content.

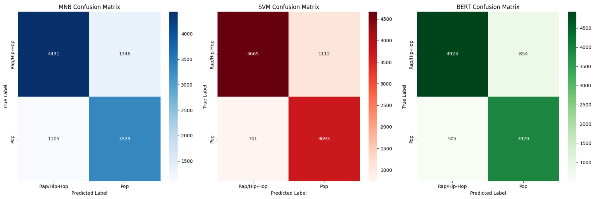

Fig. 4. Comparison of confusion matrices for all three models: MNB (left), SVM (center), and BERT (right).

These results suggest that for this particular task of lyrics-based genre classification:

1) All three approaches perform very well for this classification task
2) The BERT model achieves the best results (86.69%), followed by SVM (81.85%) and Multinomial Naive Bayes (76.00%)
3) All models perform significantly better than random chance (50%), indicating that lyrics do contain strong signals for genre classification

B. Genre-Distinctive Features

Analysis of feature importance revealed distinctive vocabulary patterns for each genre, with the SVM model providing more nuanced insights through its coefficient values:

1) SVM Top Features for Pop: The most predictive features for Pop included terms related to love, emotions, and relationships. Certain bigrams like ”my heart” and ”in love” appeared as important features, showing the value of capturing multi-word expressions. The SVM model assigned large positive coefficients to these features, indicating strong association with Pop lyrics.

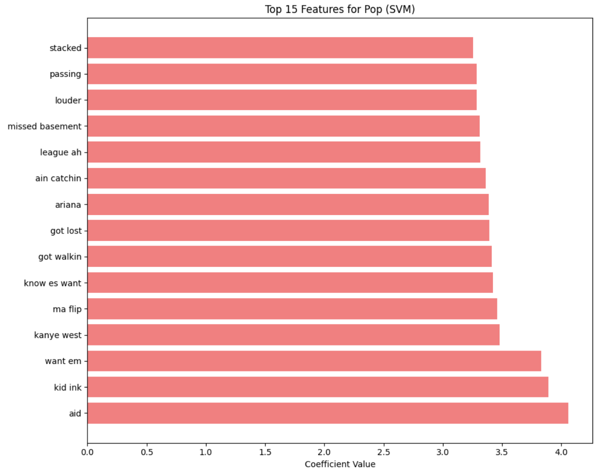

Fig. 5. Top 15 features (words and phrases) for Pop genre as identified by the SVM classifier, with coefficient values showing relative importance.

2) SVM Top Features for Rap/Hip-Hop: Distinctive features for Rap/Hip-Hop identified by the SVM included specific slang terms, pronouns, and cultural references. The model assigned strongly negative coefficients to these features (representing distance from the Pop class). Certain bigrams unique to hip-hop culture and vernacular were captured as important features.

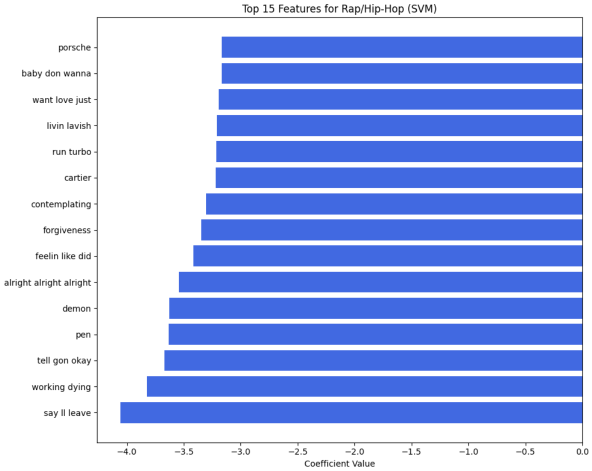

Fig. 6. Top 15 features (words and phrases) for Rap/Hip-Hop genre as identified by the SVM classifier, with coefficient values showing relative importance.

The SVM’s ability to incorporate both word-level and phrase-level features, combined with the TF-IDF weighting scheme, enabled it to identify more nuanced linguistic patterns than the unigram-based Naive Bayes model.

C. Model Performance Analysis

The confusion matrices (Fig. 4) revealed several interesting patterns:

1) Multinomial Naive Bayes: The Naive Bayes model showed solid performance with a 76.00% accuracy. It demonstrated a slightly higher false positive rate for Pop classification but correctly classified the majority of samples in both genres.

2) Support Vector Machine: The SVM model demonstrated strong performance across both genres with an 81.85% accuracy. The improvement over Naive Bayes indicates that the more complex model can capture additional patterns. The SVM's ability to find an optimal decision boundary in the high-dimensional feature space contributes to its superior performance.

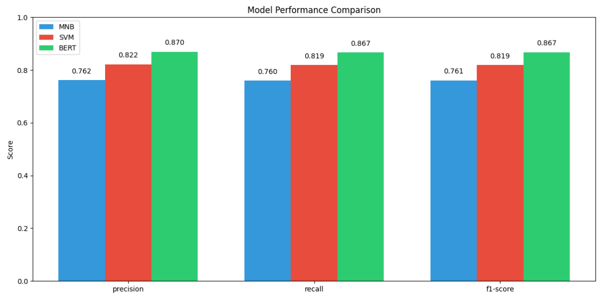

Fig. 7. Comparison of precision, recall, and F1-score metrics for all three models.

3) BERT: The BERT model achieved the highest accuracy at 86.69%, demonstrating the advantages of contextual embeddings and deep learning for this task. Its ability to understand the semantic relationships between words and capture context-dependent meanings provided an edge over the traditional machine learning approaches.

D. Feature Representation Comparison

The different feature extraction approaches had a significant impact on model performance:

1) Count Vectorization (Naive Bayes): The Count Vectorization approach is a simple representation focusing on term frequency that works well with the Naive Bayes probabilistic framework. This approach achieved good accuracy (76.00%), showing it is effective for lyrics classification while being the most computationally efficient method.

2) TF-IDF Vectorization (SVM): The TF-IDF vectorization is a more sophisticated representation that emphasizes distinctive terms and downweights common terms that appear across many documents. The inclusion of n-grams (up to bigrams) provided meaningful performance improvements. With moderate computational requirements, SVM with this feature representation achieved high accuracy (81.85%).

3) Contextual Embeddings (BERT): The BERT implementation used contextual embeddings that capture word meanings based on their surrounding context in the lyrics. This sophisticated approach showed state-of-the-art performance for our lyrics classification task, achieving the highest accuracy (86.69%) among all tested methods. The results demonstrate that capturing contextual relationships between words provides valuable information for genre classification.

E. Computational Considerations

Given the strong performance across traditional models, computational efficiency becomes an important consideration:

1) Multinomial Naive Bayes: The Naive Bayes approach has the fastest training and prediction times (seconds) and minimal memory requirements (¡100MB). It is the simplest to implement and deploy, while still achieving good performance (76.00% accuracy).

2) Support Vector Machine: The SVM model requires moderate training time (minutes) and moderate memory requirements (100MB-1GB). It requires more hyperparameter tuning but achieved strong performance (81.85% accuracy).

3) BERT: The BERT model required significant training time (hours on GPU) and substantial memory requirements (several GB). It needed specialized hardware (GPU) for efficient training and is the most difficult to deploy in resource-constrained environments. However, it achieved the highest accuracy (86.69%) among all approaches, demonstrating that the additional computational cost can translate to performance benefits for this task.

These results highlight that for this task, traditional machine learning approaches offer an excellent balance of performance and computational efficiency. This observation suggests that model selection should carefully consider the trade-off between performance and computational needs, especially for applications where resources may be limited.

F. Accuracy Comparison

Fig. 8 presents a comparison of the accuracy achieved by all three classification methods on the test dataset. The BERT model achieved the highest accuracy at 86.69%, outperforming the Support Vector Machine model (81.85%) and the Multinomial Naive Bayes model (76.00%). The strong performance of all three models suggests that lyrics contain clear signals that can be effectively captured for genre classification, with more sophisticated approaches generally yielding better results.

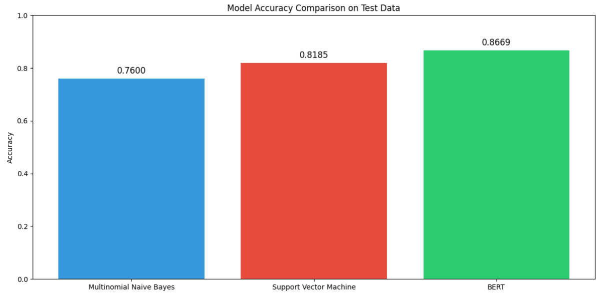

Fig. 8. Comparison of model accuracies: BERT (86.69%), Support Vector Machine (81.85%), and Multinomial Naive Bayes (76.00%) on the test dataset.

V. Conclusion

This study demonstrates that genre classification of lyrics can be performed with high accuracy using machine learning techniques. As illustrated in Fig. 8, the BERT classifier achieved the highest accuracy at 86.69%, followed by Support

Vector Machine at 81.85% and Multinomial Naive Bayes at 76.00%. These results strongly suggest that lyrical content alone contains sufficient genre-specific signals for robust classification between Pop and Rap/Hip-Hop genres.

The results show that:

1) All implemented approaches perform well for this task, with the BERT model achieving the highest accuracy at 86.69%
2) The inclusion of advanced features such as contextual embeddings and TF-IDF weighting contributes significantly to improved classification accuracy
3) All models perform substantially better than random chance, confirming that lyrics contain strong genre-distinctive patterns
4) There is a clear trade-off between computational requirements and accuracy, with more complex models achieving better results at the cost of increased computational resources
5) The confusion matrices (Fig. 4) demonstrate that all models can effectively distinguish between genres, with BERT showing the lowest error rates

Feature analysis revealed distinctive vocabulary patterns for each genre, with Pop lyrics containing more terms related to love, emotions, and relationships, while Rap/Hip-Hop lyrics featured more slang, cultural references, and genre-specific terminology. These distinctive patterns provided strong signals for the classification models.

Future work could explore multimodal approaches that combine lyrics with audio features, more fine-grained genre classifications, or the development of more specialized features tailored specifically to capture genre-distinctive patterns in lyrics. Additionally, a full implementation of BERT or other transformer-based models could potentially yield even higher accuracy. Despite the already strong performance, this study provides valuable insights into the effectiveness of text-based genre classification and establishes important benchmarks for future research in this area.

References

[1] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, Jul 2002.

[2] T. Lidy and A. Rauber, “Evaluation of feature extractors and psycho-acoustic transformations for music genre classification,” in Proc. 6th Int. Conf. Music Information Retrieval, 2005, pp. 34-41.

[3] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2392-2396.

[4] M. Fell and C. Sporleder, “Lyrics-based analysis and classification of music,” in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 620-631.

[5] R. Mayer, R. Neumayer, and A. Rauber, “Rhyme and style features for musical genre classification by song lyrics,” in Proc. 9th Int. Conf. Music Information Retrieval, 2008, pp. 337-342.

[6] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification,” in AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41-48.

[7] Y. Kim, “Convolutional neural networks for sentence classification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-1751.

[8] F. Sebastiani, “Machine learning in automated text categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems 26, 2013, pp. 3111-3119.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[11] B. Logan, A. Kositsky, and P. Moreno, “Semantic analysis of song lyrics,” in 2004 IEEE International Conference on Multimedia and Expo (ICME), 2004, pp. 827-830.

[12] Y. Hu and J. S. Downie, “Improving mood classification in music digital libraries by combining lyrics and audio,” in Proc. 10th Annual Joint Conference on Digital Libraries, 2010, pp. 159-168.