From Text to Lyrics: Evaluating the Performance of Sentiment Analysis Models on Musical Content

Dang Hoang Khang NGUYEN
Université Paris-Saclay
Orsay, France
dang-hoang-khang.nguyen@etu-upsaclay.fr

Kseniia PAVLOVA
Université Paris-Saclay
Orsay, France
kseniia.pavlova@etu-upsaclay.fr

Beining YING
Université Paris-Saclay
Orsay, France
beining.ying@etu-upsaclay.fr

***Abstract*—Sentiment analysis has been widely applied in Natural Language Processing (NLP) to analyze emotions in text. However, most sentiment classification models are trained on general textual data, making their applicability to domain-specific tasks such as musical content uncertain. This study investigates how well a sentiment classification model trained on Reddit comments (GoEmotions) can be adapted for music-related sentiment analysis. We fine-tune a pre-trained transformer-based model on GoEmotions and map its predictions to the Geneva Emotion Music Scales (GEMS-9), a framework specifically designed for categorizing emotions in music. The model's performance is evaluated on the Lyrics Emotion Dataset, assessing its effectiveness in classifying emotions within lyrical content. Experimental results indicate that while deep learning-based sentiment models exhibit reasonable performance, discrepancies exist in detecting nuanced emotions such as Solemnity and Nostalgia, which are more prevalent in lyrics than in standard text. Our findings highlight the challenges of domain adaptation in NLP models and suggest that fine-tuning on domain-specific data is necessary for improved performance in music sentiment analysis. Future work will explore multi-modal approaches, incorporating audio features alongside textual lyrics, to enhance emotion recognition accuracy.**

***Index Terms*— Sentiment Analysis, Music Emotion Recognition, NLP, GoEmotions, GEMS-9, Domain Adaptation**

I. Introduction

A. Motivation

Understanding emotions in text has been a central focus in Natural Language Processing (NLP), with sentiment analysis playing a key role in applications like customer feedback and social media monitoring. However, applying sentiment analysis models trained on general emotional text to song lyrics presents unique challenges. Lyrics often blend poetic and metaphorical language, making emotional interpretation more complex than straightforward prose. This study explores how well a sentiment analysis model trained on GoEmotions, a dataset designed for general textual emotion detection, performs when applied to song lyrics. By evaluating this model on lyrics, we aim to uncover differences in emotional expression between structured emotional text and lyrical content. Insights from this comparison could help improve sentiment analysis tools for music-related applications, such as playlist curation and listener sentiment prediction.

B. Related Works

A variety of approaches have been explored in song emotion classification, with researchers investigating both traditional machine learning models and deep learning techniques. Feature extraction from lyrics has been shown to enhance emotion classification performance [1] [2]. Various models, including Naive Bayes, HMM, SVM, clustering, and Random Forest, have been applied to lyrical and sometimes audio features to predict song emotions [3] [4] [5] [6] [7]. More recently, deep learning frameworks such as CNNs, LSTMs, and transformer-based models like BERT and ELMo have been used to improve classification accuracy [8] [9] [10] [11].

Some researchers have adopted a multi-modal approach to emotion prediction by integrating both musical and lyrical features. [12] introduced a dataset containing both music and lyrics and demonstrated improved results when incorporating both modalities. Similarly, [13] showed a significant increase in accuracy when adding lyrics to models trained on audio features alone. However, biases have been observed in emotion classification when using audio data. S [14] found that certain genres, such as heavy metal and hip-hop, were perceived to have more negative emotions compared to pop music, even when lyrics were matched. Additionally, [15] and [16] demonstrated that identical lyrics were judged differently based on the genre they were associated with, suggesting that audio features could introduce genre-related biases into emotion classification models.

Despite these advancements, research has shown that lyrics alone can be a strong predictor of song emotion. [17] used psychological feature vectors derived from lyrics and achieved human-comprehensible classification results. [3] found that lyrics-based models sometimes outperformed audio-based models for mood prediction, and that integrating both features did not always improve accuracy. Further, [18] demonstrated that lyrical features outperformed audio features in multiple mood categories, emphasizing the importance of text analysis in music emotion classification.

While many studies have explored song emotion classification through multimodal approaches and deep learning techniques, there remains a gap in research investigating the transferability of general sentiment analysis models to lyrical content. Our work differs in that it specifically evaluates a sentiment analysis model trained on general textual data (GoEmotions) and assesses its effectiveness when applied to song lyrics. This comparison allows us to better understand how emotions are expressed differently in structured emotional text versus creative lyrical writing.

The rest of this paper is structured as follows. Section II introduces our proposed solution in detail. Section III presents and analyzes the experimental results obtained from applying our methodology. Finally, we conclude this paper in Section IV and provide a link to access our work in Section V.

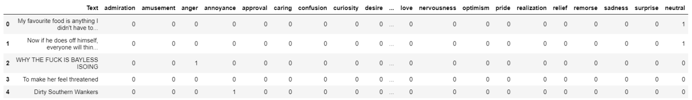
II. Method

A. Datasets

1) GoEmotions Dataset: The data used in this study includes the GoEmotions dataset, a large-scale emotion-labeled dataset consisting of over 58k carefully curated Reddit comments. This dataset has been human-annotated for 27 distinct emotion categories, along with a neutral label, making it one of the most comprehensive emotion recognition datasets available for natural language processing tasks.

- **Training set:** 43,410 examples
- **Testing set:** 5,427 examples
- **Number of Labels:** 27 emotion categories + Neutral
- **Maximum Sequence Length:** 30 tokens (in both training and evaluation datasets)

The emotion categories covered in the dataset include: Admiration, Amusement, Anger, Annoyance, Approval, Caring, Confusion, Curiosity, Desire, Disappointment, Disapproval, Disgust, Embarrassment, Excitement, Fear, Gratitude, Grief, Joy, Love, Nervousness, Optimism, Pride, Realization, Relief, Remorse, Sadness, and Surprise. Additionally, a Neutral category is assigned to texts that do not strongly convey any of the labeled emotions. Since our study aims to evaluate sentiment analysis models on song lyrics, we require a robust emotion-labeled dataset for training our models. The GoEmotions dataset is a suitable choice due to its large sample size, extensive annotation coverage, and applicability to real-world emotion recognition tasks.

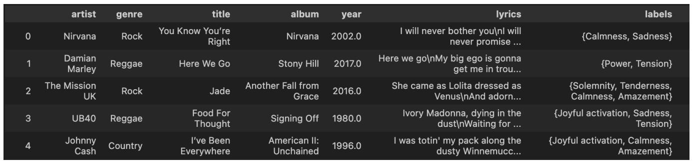


	Text	admiration	amusement	anger	annoyance	approval	caring	confusion	curiosity	desire	...	love	nervousness	optimism	pride	realization	relief	remorse	sadness	surprise	neutral
0	My favourite food is anything I didn't have to...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
1	Now if he does off himself, everyone will thin...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
2	WHY THE FUCK IS BAYLESS ISOING	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	To make her feel threatened	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Dirty Southern Wankers	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Fig. 1: Example format of the GoEmotions dataset. Each text is annotated with one or more emotion labels.

2) Lyrics Emotion Dataset: In addition to the GoEmotions dataset, we also use a dataset of song lyrics for evaluation purposes. This dataset consists of 1,160 song lyrics hand-annotated using 9 categories of the Geneva Emotional Music Scales (GEMS). The dataset is mainly used for testing to examine if an emotion classification model trained on general text can effectively classify emotions in song lyrics.

- **Total Number of Examples:** 1,160 examples
- **Number of Labels:** 9 emotion categories
- **Emotion Categories:** Amazement, Calmness, Joyful activation, Nostalgia, Power, Sadness, Solemnity, Tenderness, Tension



	artist	genre	title	album	year	lyrics	labels
0	Nirvana	Rock	You Know You're Right	Nirvana	2002.0	I will never bother you\nI will never promise ...	{Calmness, Sadness}
1	Damian Marley	Reggae	Here We Go	Stony Hill	2017.0	Here we go\nMy big ego is gonna get me in trou...	{Power, Tension}
2	The Mission UK	Rock	Jade	Another Fall from Grace	2016.0	She came as Lolita dressed as Venus\nAnd adorn...	{Solemnity, Tenderness, Calmness, Amazement}
3	UB40	Reggae	Food For Thought	Signing Off	1980.0	Ivory Madonna, dying in the dust\nWaiting for ...	{Joyful activation, Sadness, Tension}
4	Johnny Cash	Country	I've Been Everywhere	American II: Unchained	1996.0	I was totin' my pack along the dusty Winnemucc...	{Joyful activation, Calmness, Amazement}

Fig. 2: Example format of the Lyrics Emotion dataset. Each text is annotated with one or more emotion labels.

The Lyrics Emotion dataset is crucial in this study as it allows us to examine the effectiveness of models trained on the GoEmotions dataset when applied to a domain with different textual characteristics. Lyrics often contain metaphorical and poetic expressions, which may differ significantly from the conversational text of Reddit comments. By comparing model performance on both GoEmotions and the Lyrics Emotion dataset, we aim to understand the extent to which sentiment analysis models can generalize across different textual domains. ‘

B. Exploratory Data Analysis

1) GoEmotions Dataset: To better understand the distribution of emotions in the datasets, we conducted an exploratory data analysis (EDA). The first analysis examines the frequency distribution of emotions within the GoEmotions dataset. The results, visualized in Figure 3, show that the neutral category is the most prevalent, followed by admiration, approval, and gratitude. Emotions such as grief and realization appear much less frequently, indicating an imbalance in the dataset.

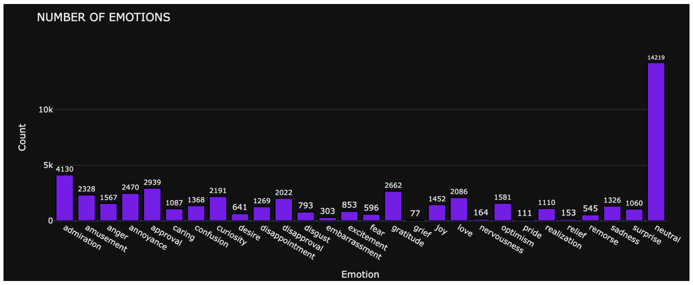


Fig. 3: The frequency of each emotion labels in the GoEmotions dataset

In addition, we generated word clouds for each emotion category to highlight the most frequently associated words. Figure 4 presents these visualizations, where words such as ”great” and ”awesome” are prominent in admiration, whereas words like ”angry” and ”hate” dominate the anger category. This helps illustrate how different emotions are expressed in text.

2) Lyrics Emotion Dataset: The emotion distribution in the Lyrics Emotion dataset, visualized in Figure 5, reveals a different dominant emotion compared to GoEmotions. Sadness is the most prevalent emotion in song lyrics, followed by

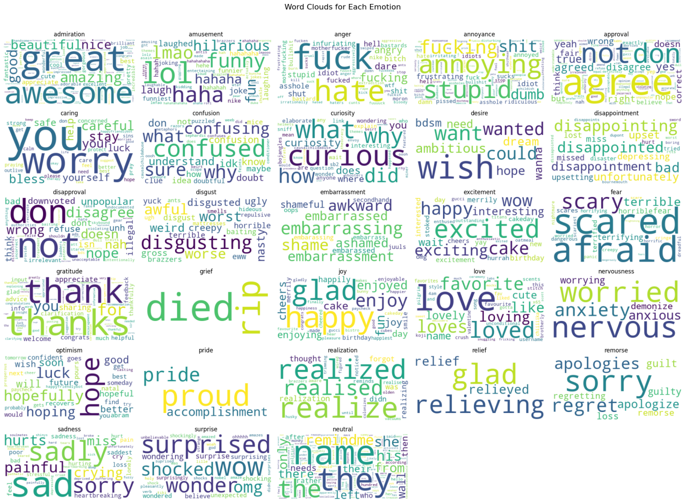


Fig. 4: The wordclouds of each emotion labels in the GoEmotions dataset

tension and tenderness, whereas GoEmotions is dominated by neutral expressions. This discrepancy highlights the challenge of applying a sentiment analysis model trained on general text to lyrics, as the emotional structure of lyrics differs significantly from conversational Reddit comments.

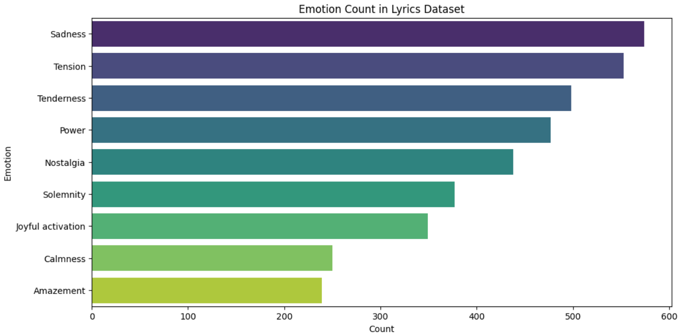


Fig. 5: Emotion frequency distribution in the Lyrics Emotion dataset.

We also created word clouds for each emotion category in the Lyrics Emotion dataset (Figure 6). Unlike GoEmotions, where common conversational words dominate, lyrics exhibit more poetic and metaphorical word usage. Words like ”you,” ”the,” and ”me” appear frequently across emotions, reflecting the narrative and personal nature of song lyrics. Understanding these differences is crucial for adapting models trained on general text to effectively capture sentiment in lyrics.

C. Data Preprocessing

To ensure the text data is structured, clean, and ready for model training, we applied a series of preprocessing steps. These steps help remove noise, standardize text formats, and improve the quality of textual representations for sentiment analysis models.

The preprocessing pipeline consists of the following steps:

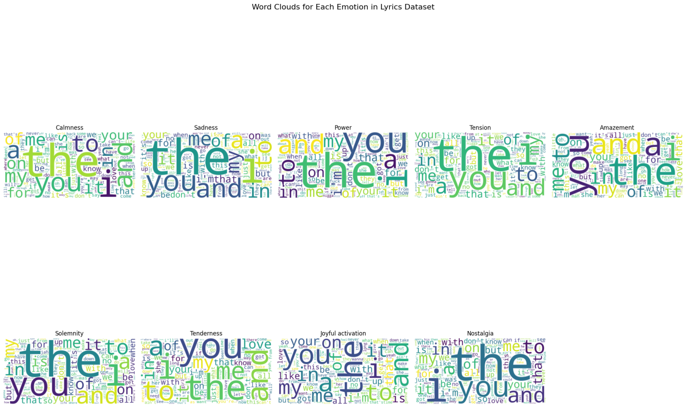


Fig. 6: Word clouds for each emotion category in the Lyrics Emotion dataset.

- **Text Cleaning:** We remove special characters, extra whitespaces, and lowercase all text to maintain consistency across datasets.
- **Tokenization:** The text is split into individual words or subwords using word-based or subword-based tokenization techniques, depending on the model requirements.
- **Stopword Removal:** Common but uninformative words (e.g., ”the,” ”and,” ”is”) are eliminated to focus on meaningful content.
- **Stemming:** Words are reduced to their base or root forms to consolidate different morphological variations.
- **Handling Emojis and Character Repetition:** Emojis are standardized or replaced with corresponding text labels. Additionally, elongated words (e.g., ”coooool”) are shortened to their base form.
- **Entity and Special Token Replacement:** Usernames, currency symbols, and other special tokens are replaced with generic placeholders (e.g., `u/username` → `<user>`, 4 → `<price>`).
- **Emoticon Replacement:** Common emoticons (e.g., `:))` are mapped to their textual descriptions (e.g., `<smile_face>`).
- **Expanding Contractions:** Contractions such as `you're` and `isn't` are expanded to `you are` and `is not`, respectively, to maintain textual clarity.
- **Spelling Correction:** Common spelling errors are corrected, standardizing variations between American and British English (e.g., `colour` → `color`).

D. Model Selection

To evaluate sentiment analysis performance, we developed two models for comparison. o handle the multi-label nature of the emotion classification task, we use a **One-vs-Rest (OvR)** classification strategy, where a separate binary SVM classifier is trained for each emotion category. This allows the model to assign multiple emotion labels to a single text.

For classification, we employ a **Support Vector Machine (SVM)** with a **linear kernel**. The linear kernel is particularly effective for high-dimensional text data, ensuring efficient computation and robust classification performance. The model is trained using the default hinge loss, optimizing the margin between emotion classes.. The baseline model employs Term Frequency-Inverse Document Frequency (TF-IDF) for feature representation, while the second model utilizes RoBERTa for word embeddings. This distinction allows us to examine the impact of different text representations on sentiment classification performance.

E. Emotion Mapping

Evaluating the Lyrics Emotion dataset presents a challenge as its emotion labels do not directly match those in the GoEmotions dataset. While GoEmotions contains 27 fine-grained emotion labels plus a neutral category, the Lyrics Emotion dataset uses 9 broader categories based on the Geneva Emotional Music Scales (GEMS-9). To enable cross-dataset evaluation, we perform an emotion mapping process.

The mapping process assigns multiple fine-grained emotions from GoEmotions to a single high-level category in GEMS-9. This allows for meaningful comparisons between the datasets. After mapping, the classifier model predicts confidence scores for each of the original 28 GoEmotions classes. These scores are then summed within each corresponding GEMS-9 category to produce aggregated confidence values for the broader emotion classes.

Since the Lyrics Emotion dataset allows multiple emotion labels per song, we use a ranking approach: the model selects the top predicted labels with the highest confidence scores. This ensures that the evaluation process aligns with the multi-label nature of song lyrics and allows for direct comparison between the predicted and true labels.

Table I provides an example of the emotion mapping strategy:

GEMS-9 Emotion	**Mapped GoEmotions Categories**
Nostalgia	admiration, sadness, love, realization
Tension	fear, nervousness, annoyance, anger
Calmness	relief, approval, gratitude
Joyful Activation	joy, excitement, amusement
Amazement	surprise, excitement, realization
Tenderness	caring, love, admiration
Power	pride
Sadness	sadness, grief, disappointment, remorse
Solemnity	disapproval, sadness, realization

TABLE I: Emotion Mapping from GoEmotions to GEMS-9 Categories.

III. Results

This section presents the results of the sentiment analysis models on both the GoEmotions and Lyrics Emotion datasets. We evaluate model performance using standard classification metrics, including precision, recall, F1-score, Jaccard similarity score, and Hamming loss. These metrics are chosen to provide a comprehensive evaluation of the models' effectiveness in multi-label classification. Additionally, we analyze the differences in performance across datasets to assess the generalization capability of the models.

A. Evaluation Metrics

Given that our task involves multi-label emotion classification, traditional accuracy metrics are insufficient. Instead, we focus on the following evaluation metrics:

Precision and Recall: Precision measures the proportion of correctly predicted positive labels out of all predicted positive labels, while recall measures the proportion of correctly predicted positive labels out of all actual positive labels. These are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

where represents true positives, represents false positives, and represents false negatives.

F1-score (Micro and Macro): The F1-score is a harmonic mean of precision and recall, making it an essential metric for imbalanced datasets. The micro-averaged F1-score aggregates contributions from all classes to compute the average metric, while the macro-averaged F1-score gives equal weight to all classes. The formulas are as follows:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Jaccard Similarity Score: Measures the similarity between the predicted and true label sets. It is defined as the intersection over the union of predicted and true labels:

$$\text{Jaccard Similarity} = \frac{|Y_{true} \cap Y_{pred}|}{|Y_{true} \cup Y_{pred}|} \tag{4}$$

Hamming Loss: Represents the fraction of labels that are incorrectly predicted. It is useful for multi-label classification as it penalizes both false positives and false negatives:

$$\text{Hamming Loss} = \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} I(y_{ij} \neq \hat{y}_{ij}) \tag{5}$$

where is the total number of samples, is the number of labels, and is an indicator function that returns 1 when the predicted label differs from the true label.

B. Performance on GoEmotions Dataset

The first evaluation examines how well each model performs on the GoEmotions dataset. Given that this dataset was used for training, the results indicate how effectively the models learn emotion classification in a general textual context.

Tables II and III present the performance of the baseline **TF-IDF + SVM** model and the advanced **RoBERTa-based** model on the GoEmotions test set.

Metric	Precision	Recall	F1-score
Macro Avg	0.68	0.25	0.32
Weighted Avg	0.72	0.35	0.42
Samples Avg	0.39	0.37	0.38

TABLE II: Performance of SVM (TF-IDF) on the GoEmotions dataset.

Metric	Precision	Recall	F1-score
Macro Avg	0.57	0.39	0.45
Weighted Avg	0.66	0.51	0.56
Samples Avg	0.57	0.54	0.55

TABLE III: Performance of RoBERTa on the GoEmotions dataset.

The results indicate that SVM achieves higher precision but lower recall, suggesting that it is more conservative in its predictions, potentially avoiding false positives but missing many correct emotion labels. In contrast, RoBERTa demonstrates higher recall and F1-score, meaning it captures a broader range of emotions but at the cost of more false positives.

Additionally, the macro-averaged recall for SVM is particularly low (0.25), which highlights its struggle in detecting minority emotion classes (e.g: relief, grief). Meanwhile, RoBERTa, with a macro recall of 0.39, performs significantly better at capturing less frequent emotions. This suggests that deep learning approaches such as RoBERTa generalize better for nuanced emotional expressions compared to traditional TF-IDF-based models.

Overall, these findings suggest that while SVM is effective in precision-oriented tasks, RoBERTa's improved recall makes it more suitable for capturing a broader emotional spectrum, particularly in datasets where emotions may overlap or be expressed in varying linguistic styles.

C. Performance on Lyrics Emotions Dataset

Table IV and Table V present the evaluation results of the SVM and RoBERTa models on the Lyrics Emotion dataset. The performance metrics include Jaccard similarity score, F1-score (Micro and Macro), and Hamming loss.

Metric	SVM
Jaccard Similarity Score	0.2935
F1-Score (Micro)	0.4215
F1-Score (Macro)	0.2895
Hamming Loss	0.3820

TABLE IV: Performance of SVM on the Lyrics Emotion dataset.

Metric	RoBERTa
Jaccard Similarity Score	0.3509
F1-Score (Micro)	0.4829
F1-Score (Macro)	0.3767
Hamming Loss	0.3283

TABLE V: Performance of RoBERTa on the Lyrics Emotion dataset.

The results show that RoBERTa outperforms SVM in all key performance metrics. Specifically, RoBERTa achieves a higher Jaccard similarity score and F1-scores, indicating a better ability to assign relevant emotion labels to song lyrics. Notably, the macro F1-score improves significantly from 0.2895 (SVM) to 0.3767 (RoBERTa), suggesting that RoBERTa captures a more diverse range of emotions.

Additionally, RoBERTa achieves a lower Hamming Loss (0.3283 vs. 0.3820), meaning that it makes fewer incorrect label assignments per sample. This highlights its superior handling of multi-label classification in the lyrics dataset.

These findings suggest that deep learning approaches, such as RoBERTa, are better suited for emotion classification in song lyrics compared to traditional methods like TF-IDF + SVM, which struggle to capture nuanced emotions due to their limited contextual understanding

D. Discussion

The results highlight key differences in model performance across datasets. While RoBERTa demonstrates superior performance compared to SVM, both models struggle with domain adaptation from Reddit-based text (GoEmotions) to song lyrics. This suggests that additional fine-tuning on music-related text could improve performance.

Error analysis reveals that SVM often misclassifies minority emotion classes, as indicated by its low macro F1-score. RoBERTa performs better in capturing nuanced emotions, yet it still struggles with overlapping categories, particularly between sadness and nostalgia. Further investigation into contextual embeddings and transfer learning techniques may help mitigate these issues.

IV. Conclusion

This study evaluated the performance of sentiment analysis models trained on GoEmotions and tested on song lyrics. The results show that RoBERTa outperforms the baseline SVM model, particularly in recall and F1-score, highlighting its better generalization to lyrics-based sentiment classification. However, both models exhibit limitations in adapting to the musical domain without fine-tuning. The emotion mapping strategy enabled cross-domain evaluation, but discrepancies between fine-grained and high-level emotion categories remain a challenge. Future work should explore fine-tuning deep learning models on music-related text to enhance sentiment classification accuracy.

Beyond the technical aspects, this study provided hands-on experience with NLP models for sentiment classification, offering insights into the challenges of applying a generalized model to a specific domain. The findings emphasize the importance of fine-tuning deep learning models for better adaptation to domain-specific tasks. These insights contribute to a broader understanding of the strengths and weaknesses of current NLP methodologies in real-world applications.

V. Data and Code Availability

The notebook and data is available at Github: https://github.com/KsenaPav/NLP_final_project

VI. Contribution

- **Evaluation of NLP sentiment models on music-related content:** We systematically assess the applicability of sentiment analysis models trained on general textual data to the domain of musical lyrics. Unlike prior studies that focus on multimodal emotion recognition, our work provides a dedicated analysis of pre-trained NLP models in the musical domain.
- **Mapping general sentiment labels to music-specific emotions:** We introduce a novel mapping framework from GoEmotions' general sentiment labels to the Geneva Emotion Music Scales (GEMS-9), bridging the gap between generic emotion datasets and domain-specific emotion classification in music.
- **Domain adaptation challenges in sentiment analysis:** Our findings highlight the difficulty of adapting sentiment analysis models to lyrical text, revealing that nuanced emotions such as *Solemnity* and *Nostalgia* are more prevalent in song lyrics than in general textual data.
- **Insights for future research:** We emphasize the need for multimodal approaches integrating both textual and audio features to improve music sentiment analysis, paving the way for more robust and context-aware emotion recognition in music-related applications.

References

[1] Y. X. B. C. W. S. L. Z. H He, J Jin, “Language feature mining for music emotion classification via supervised learning from lyrics,” *Advances in Computation and Intelligence: Third International Symposium*, 2008.

[2] D. Y. Y. W. X Wang, X Chen, “Music emotion classification of chinese songs based on lyrics using tf* idf and rhyme.” *ISMIR*, 2011.

[3] A. E. X Hu, JS Downie, “Lyric text mining in music mood classification,” *American music*, 2009.

[4] H. K. M Kim, “Lyrics-based emotion classification using feature selection by partial syntactic analysis,” *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, 2011.

[5] K. K. R. D. A Jamdar, J Abraham, “Emotion analysis of songs based on lyrical and audio features,” *arXiv preprint*, 2015.

[6] S. W. Y An, S Sun, “Naive bayes classifiers for music emotion classification based on lyrics,” *2017 IEEE/ACIS 16th International Conference on Computer and Information Science*, 2017.

[7] C. F. FH Rachman, R Sarno, “”music emotion classification based on lyrics-audio using corpus based emotion,” *International Journal of Electrical Computer Engineering*, 2018.

[8] F. P. J. R.-L. M. M. R Delbouys, R Hennequin, “Music mood detection based on audio and lyrics with deep neural net,” *arXiv preprint*, 2018.

[9] Y. W. J Abdillah, I Asror, “Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting,” *Rekayasa Sistem Dan Teknologi Informasi*, 2020.

[10] S. O. M. T. L Parisi, S Francia, “Exploiting synchronized lyrics and vocal features for music emotion detection,” *arXiv preprint*, 2019.

[11] Z. T. G Liu, “Research on multi-modal music emotion classification based on audio and lyirc,” *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference*, 2020.

[12] A. B. C Strapparava, R Mihalcea, “A parallel corpus of music and lyrics annotated with emotions,” *LREC, 2012*, 2012.

[13] H. C. I. L.-Y. H. H. C. YH Yang, YC Lin, “Toward multi-modal music emotion classification,” *Advances in Multimedia Information Processing-PCM*, 2008.

[14] E. S. M Susino, “Negative emotion responses to heavy-metal and hip-hop music with positive lyrics,” *Empirical Musicology Review*, 2019.

[15] N. S. A Dunbar, CE Kubrin, “The threatening nature of “rap” music,” *Psychology, Public Policy, and Law*, 2016.

[16] C. Fried, “Who’s afraid of rap: Differential reactions to music lyrics 1,” *Journal of Applied Social Psychology*, 1999.

[17] W. L. D Yang, “Music emotion identification from lyrics,” *2009 11th IEEE International Symposium on Multimedia*, 2009.

[18] J. D. X Hu, “When lyrics outperform audio for music mood classification: A feature analysis,” *Ismir*, 2010.