Mood Prediction from Song Lyrics

Anna Basil Jose (annjo768)
732A81 - Text Mining
annjo768@student.liu.se

Abstract

In the era of digital music streaming, accurately determining the mood of songs through their lyrics has become crucial for enhancing user experience. This project delves into the use of Natural Language Processing (NLP) techniques for mood prediction in song lyrics, employing various baseline models such as Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest Classifier with both TF-IDF and Count Vectorization methods. Additionally, advanced models like DistilBERT and an LSTM network with GloVe embeddings were tested. The comparative analysis focused on metrics such as accuracy, precision, recall, F1 scores, and confusion matrices to evaluate model performance. Among the models, DistilBERT emerged as the most effective, demonstrating superior accuracy and F1 score, suggesting its transformer architecture's enhanced capability to understand contextual nuances in language over traditional and LSTM-based models. This finding highlights the potential of transformer-based models in processing complex language tasks, offering a promising avenue for improving mood prediction systems based on lyrics.

1 Introduction

"Life seems to go on without effort when I am filled with music." – George Eliot. Music is a unique medium that crosses cultural barriers and profoundly speaks to the human soul. According to (statistical Portal, 2023), the monthly active users of a famous music streaming service, Spotify, increased nearly ten times from 68 Million in 2015 to 608 Million by the end of 2023. In 2023, each Spotify user streamed approximately 3696 minutes on average of music according to a calculation done by (Eliaçık, 2023). Human beings listen to music for several reasons - to regulate mood, as self-therapy, and to boost productivity (Sloboda and O'neill, 2001), (Lesiuk, 2005). People often choose music with emotion congruent to their situation. Studies also revealed that music can alter the mood of a person depending upon the mood of the music they listen to (Thoma et al., 2012), (Sloboda and O'neill, 2001). Researchers have for so long identified that the lyrics of a song have a significant impact on determining the mood associated with the song.

Numerous literature use different deep learning (DL) techniques to get the mood of the song from the lyrics (Tzanetakis), (Juthi et al., 2020). While DL and machine learning (ML) models provide an underlying framework and methods for the prediction, Natural Language Processing (NLP) helps apply these to the natural language as both DL and ML techniques naturally do not deal with language. This project will explore various baseline models, including the Naive Bayes Classifier, Support Vector Machine, and Random Forest Classifier, that use two different vectorization techniques - Count Vectorization and Term Frequency - Inverse Document Frequency (TD-IDF) Vectorization - to establish initial predictions. It will then advance to more complex NLP models such as LSTM with GloVe embeddings and DistilBERT, aiming to achieve a more accurate prediction of song moods. This approach gives a comprehensive comparison of a few traditional and newer techniques in the context of mood prediction by lyric analysis.

2 Theory

2.1 LSTM with GloVe embeddings

Although a typical Recurrent Neural Network (RNN) may predict the next word in a text sequence by considering all of its prior words, training an RNN can be difficult. In order to overcome the drawbacks of RNN as well as the vanishing gradient issue, Long Short-Term Memory (LSTM), a distinct kind of RNN, was first introduced by (Hochreiter and Schmidhuber, 1997). LSTM has

been used for several applications where the sequence of words and their context are crucial for understanding sentiment as learning long-term dependencies in data sequences makes them particularly effective for processing and making predictions based on text.

Global Vectors for Word Representation (GloVe) embeddings provide a powerful way to encode words into dense vectors that capture semantic relationships between words based on their co-occurrence in a large text corpora (Pennington et al., 2014). GloVe constructs embeddings by analyzing word co-occurrences across a vast corpus, creating a global co-occurrence matrix. This technique ensures that the embeddings encapsulate the semantic relationships between words based on their contextual associations, essential for grasping the nuanced meanings within lyrics.

Combining LSTM with GloVe embeddings allows the mood prediction model to leverage GloVe's pre-trained word representations to understand the semantics of the lyrics better and use LSTM's sequence modeling capabilities to understand the context and flow of the text. This combination is powerful because it enables the model to capture both the meaning of individual words and their sequence, which is critical for accurately determining the mood conveyed by a piece of text.

2.2 DistilBERT

DistilBERT is a compact, efficient version of the BERT (Bidirectional Encoder Representations from Transformers) model, designed to balance between processing speed and comprehension capability efficiently (Sanh et al., 2020). It is achieved through knowledge distillation, a process that trains a smaller model (DistilBERT) to mimic the behavior of a larger, more complex model (BERT). Despite being 40% smaller, DistilBERT preserves about 95% of BERT's language processing abilities while offering a 60% increase in speed.

Due to its transformer architecture, DistilBERT is adept at capturing both the context and semantic depth of words in sequences, an essential feature for accurately identifying the mood conveyed in song lyrics. This model's streamlined size and enhanced speed make it an exceptional choice for analyzing extensive text data efficiently, without significantly sacrificing accuracy. This capability is invaluable for mood prediction in lyrics, where linguistic analysis is needed.

3 Data

This project uses a lyric dataset created by the developers of the project - *Music Mood Prediction* (Manojit, 2023). The dataset contains two columns: Lyrics and Mood. The moods are encoded as Happy - 1, Sad - 2, Angry - 3, and Relaxed - 4. The dataset is created by the following steps. A dataframe with the following columns: File name (name of the h5 file), Artist Name, Song Title, and Lyrics (Empty now) is created from the Million Song Dataset. Using the Artist Name and Song Title, lyrics for all the songs are fetched using the PyLyrics package, which uses LyricWikia API. Songs that are not in English are removed from the dataset. Last.FM API is used to extract the tags for the remaining songs. The mood of each song is assigned by correlating these tags to a set of words associated with each mood. Yet another dataset obtained from the paper (Xue et al., 2015) is appended to the created dataset. The authors extended this dataset by translating a few Hindi songs to English using Google Translate API and manually labeling them. The final dataset that the project (Manojit, 2023) offers has 1716 lyrics-mood pairs with 547 rows belonging to the Happy class, 586 entries belonging to the Sad class, 218 songs belonging to the Angry class, and 365 rows belonging to the Relaxed class.

As seen above from the numbers, the dataset exhibits a significant imbalance, with the number of entries in the highest class being more than twice that of the least represented class. To create a more balanced dataset, we use a data augmentation technique that makes use of MarianMTModel. The lyrics of the songs belonging to the minority class are converted to another language, here France, and back. This gives a new set of songs with synonyms lyrics. This is done for songs belonging to the Angry class in our dataset to augment 100 new entries. The dataset used in this project hence has 1816 lyrics-mood pairs with 547 rows belonging to the Happy class, 586 entries belonging to the Sad class, 318 songs belonging to the Angry class, and 365 rows belonging to the Relaxed class.

4 Method

This section outlines the models implemented in this project to predict song moods based on their lyrics. The approach is divided into two phases: training baseline models to establish benchmarks and exploring advanced models to enhance predic-

tion accuracy.

4.1 Baseline models

The initial phase involved experimenting with four baseline models:

- Multinomial Naive Bayes,
- Support Vector Machine (SVM),
- Logistic Regression
- Random Forest Classifier.

These models were chosen due to their popularity for classification tasks.

4.1.1 Text Vectorization

Prior to model training, the lyrics data was preprocessed through vectorization, converting text data into numerical format. Two vectorization techniques were applied:

Count Vectorization: Transforms the text data into vectors based on the number of word occurrences within the document.

TF-IDF Vectorization: Converts text to vectors by reflecting how important a word is to a document in a collection or corpus by taking into consideration both the Term Frequency and Inverse Document Frequency.

4.1.2 Prediction and Evaluation

The trained models were used to predict the mood of songs in the test set. Performance metrics such as accuracy, precision, recall, and F1 score were calculated to evaluate each model's effectiveness. A confusion matrix was also generated to visually assess model performance across different mood categories.

4.2 Advanced Models

To improve upon the baseline models, two advanced NLP models were explored:

- LSTM with GloVe Embeddings
- DistilBERT

4.2.1 Implementation Details

LSTM with GloVe Embeddings: The pre-trained GloVe vector used in this project is **glove.6B.100d**. It contains a set of pre-trained word vectors, developed using data from Wikipedia 2014 and Gigaword 5, containing 6 billion tokens and a vocabulary of 400,000 words, represented as 100-dimensional vectors.

DistilBERT: The **distilbert-base-uncased** pretrained model is used in this project to predict the mood of the songs from their lyrics.

4.2.2 Prediction and Evaluation

Evaluation after model training is conducted as similar to that of the baseline models. Metrics including accuracy, precision, recall, and F1 score were calculated to compare the prediction. To get a visual output of the prediction, a confusion matrix was generated for each model.

5 Results

5.1 Result of Baseline Models

The results of the baseline models with both TF-IDF and Count Vectorization can be seen in Table 1 and Table 2.

The results indicate that models trained with TF-IDF vectorization generally outperform those trained with Count Vectorization on this mood prediction task, suggesting TF-IDF's effectiveness in emphasizing informative words while mitigating the impact of common but less informative words across documents.

Model	Accuracy	Precision	Recall	F1Score
RF	0.51	0.51	0.51	0.49
LR	0.51	0.51	0.51	0.49
SVM	0.49	0.56	0.49	0.46
NB	0.39	0.69	0.39	0.30

Table 1: Comparison of baseline models with TF-IDF Vectorization.

Model	Accuracy	Precision	Recall	F1Score
RF	0.47	0.47	0.47	0.47
LR	0.47	0.47	0.47	0.47
SVM	0.41	0.48	0.41	0.35
NB	0.48	0.50	0.48	0.47

Table 2: Comparison of baseline models with Count Vectorization.

The Logistic Regression (LR) and Random Forest models achieve identical performance metrics with both vectorization techniques, highlighting their robustness and consistency across different feature representations.

Under TF-IDF, the NB model shows notably high precision but significantly lower accuracy and F1 score, indicating a tendency to make correct predictions when it does classify a sample as positive

but struggles with overall classification balance. The SVM model exhibits a solid balance between accuracy and precision, suggesting its capability to effectively leverage the weighted features provided by TF-IDF. The LR and Random Forest models, sharing identical metrics, demonstrate moderate effectiveness.

With Count Vectorization, the NB model shows a slight improvement in accuracy over TF-IDF Vectorization but with a decrease in precision, illustrating a trade-off between generalizing across the dataset and making precise classifications. The SVM model's performance drops significantly, indicating challenges in handling the high-dimensional, sparse feature sets without the normalization benefits of TF-IDF. LR and Random Forest models display a slight dip in performance metrics compared to TF-IDF, reaffirming the complexity and nuances involved in mood prediction that TF-IDF's nuanced feature weighting helps to somewhat mitigate.

5.2 Result of Advanced models

Results of both DistilBERT and LSTM with GloVe is shown in the table 3.

Model	Accuracy	Precision	Recall	F1
1	0.42	0.41	0.42	0.38
2	0.50	0.52	0.50	0.50

Table 3: Comparison of advanced models - Model 1:LSTM with GloVe Embeddings, Model 2: DistilBERT.

The evaluation metrics reveal distinct performance characteristics of the DistilBERT model and the LSTM network with GloVe embeddings in the context of a mood prediction task. DistilBERT achieved an accuracy of approximately 50.3%, with precision, recall, and F1 scores slightly above or around the 50% mark, indicating a balanced ability to correctly identify the mood of text inputs while maintaining a relatively equal balance between precision and recall. In contrast, the performance of the LSTM model shows a lower accuracy of around 42.1% and an F1 score of approximately 38.2%, suggesting it struggled more with the mood prediction task. This suggests DistilBERT's superior ability to contextualize and understand the nuances of text for mood prediction, benefiting from the transformer architecture's efficient handling of sequence data compared to LSTM approaches.

6 Discussion

In analyzing the results, it becomes evident that transformer-based models like DistilBERT offer a compelling advantage in natural language processing tasks, including mood prediction from text, over traditional machine learning models and even more sophisticated neural networks like LSTM with GloVe embeddings. The superior performance of DistilBERT can be attributed to its ability to capture contextual dependencies and nuances in language through its transformer architecture. This is particularly beneficial for mood prediction, where the context and subtleties of language play crucial roles.

However, this work is not without limitations. The dependency on pre-trained embeddings and models requires substantial computational resources, making accessibility an issue for some researchers. Furthermore, the black-box nature of deep learning models, especially transformers, poses challenges in interpretability and understanding model decisions.

7 Conclusion

This project undertook the task of mood prediction from lyrics of songs, leveraging both traditional machine learning models with TF-IDF and Count Vectorization techniques, and advanced NLP models utilizing pre-trained embeddings and architectures. The analysis revealed that DistilBERT, a transformer-based model, provided the most accurate predictions compared to the other models, underscoring the effectiveness of transformers in handling complex language tasks.

The comparison of different models highlights the trade-offs between model complexity and performance, offering insights into the practical application of advanced NLP models. Future work could explore reducing the computational costs of transformer models, improving model interpretability, and expanding the dataset to cover a broader range of moods and linguistic expressions as well as balancing the dataset, further enhancing the robustness and applicability of mood prediction models.

References

Eray Eliaçık. 2023. Wrapped 2023: What is the average spotify listening time? https://dataconomy.com/2023/11/30/what-is-average-spotify-listening-time/.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jannatul Humayra Juthi, Anthony Gomes, Touhid Bhuiyan, and Imran Mahmud. 2020. Music emotion recognition with the extraction of audio features using machine learning approaches. In *Proceedings of ICETIT 2019: Emerging Trends in Information Technology*, pages 318–329. Springer.

Teresa Lesiuk. 2005. The effect of music listening on work performance. *Psychology of music*, 33(2):173–191.

Manojit. 2023. MusicMoodPrediction. https://github.com/manojit32/MusicMoodPrediction.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

John A Sloboda and Susan A O'neill. 2001. Emotions in everyday listening to music. *Music and emotion: Theory and research*, 8:415–429.

The statistical Portal. 2023. Spotify maus worldwide 2023.

Myriam V Thoma, Stefan Ryf, Changiz Mohiyeddini, Ulrike Ehlert, and Urs M Nater. 2012. Emotion regulation through listening to music in everyday situations. *Cognition & emotion*, 26(3):550–560.

George Tzanetakis. Marsyas submissions to mirex 2009.

Hao Xue, Like Xue, and Feng Su. 2015. Multimodal music mood classification by fusion of audio and lyrics. In *MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II 21*, pages 26–37. Springer.