Text Mining and Social Media Mining

Tweets to Tunes: Unraveling Depressive Content in Artistic Expression

Nurdan Beşli, 457945

Maciej Lorens, 419763

Huseyin Polat, 437969

February, 2024

Table of Content

1. Introduction

In today's digital era, social media serves as a significant platform for human expression. This project specifically employs the emotions conveyed through social media, a space where individuals share various aspects of their lives, including positive experiences, thoughts, and challenges. The focus is on Twitter, a widely utilized social media platform predominantly centered around textual content. Our journey begins with the analysis of tweet data labeled with suicidal intention, seeking to understand the patterns associated with expressions indicative of depressive states. At the core of our approach is a Convolutional Neural Network (CNN) paired with a Long Short-Term Memory (LSTM) architecture, providing a sophisticated tool for identifying and categorizing depressive content.

The CNN-LSTM model, tuned to detect nuanced markers of depressive content, sets the stage for the next phase. Second, we extend our exploration into the world of music, employing the trained model to analyze and quantify depressive expressions in the lyrics of various artists. This bridge between social media data and song lyrics offers a unique lens through which we can understand how individuals communicate and share their emotional struggles.

As technology continues to integrate with human experiences, our project sits at the intersection of computational analysis and emotional exploration. By harnessing the power of deep learning models, our aim is to discern patterns in social media content and extend our understanding of emotional expression into the realm of art.

2. Data

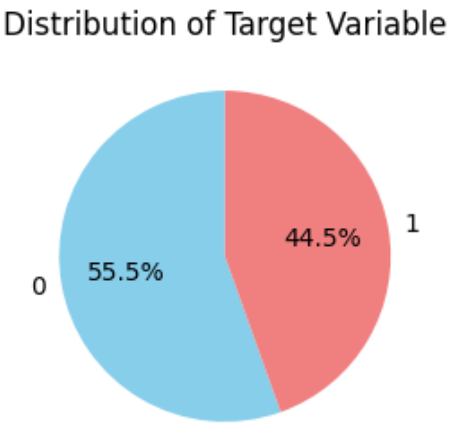
2.1. Dataset

The data for this analysis originates from different sources.

For the training of CNN-LSTM model, a combination of labelled suicidal intention tweet datasets from Kaggle and Github are utilized. Below is the list of data sources with the links embedded.

1. Twitter suicidal intention dataset
2. Suicidal Tweet Detection Dataset
3. Suicidal Phrases

A notable observation is that there are overlapping tweets across the datasets. To ensure the training dataset is devoid of duplicate entries, only a distinct set of tweets is retained. Among these, 44.5% are labeled with suicide intention, while 55.5% are categorized as neutral.



For the subsequent phase of this project, which involves analyzing and quantifying depressive expressions in the lyrics of different artists, a comprehensive lyrics dataset from Kaggle is employed. The data source and their respective link are provided below.

4. Song Lyrics Dataset

Table below represents data volume in all datasets mentioned above.

Dataset	#of Tweets	#of Unique Tweets	# of Artists	# of Albums	# of Songs
Twitter suicidal intention dataset	9,119	10,824	-	-	-
Suicidal Tweet Detection Dataset	1,599		-	-	-
Suicidal Phrases	1,787		-	-	-
Song Lyrics Dataset	-	-	21	557	5,981

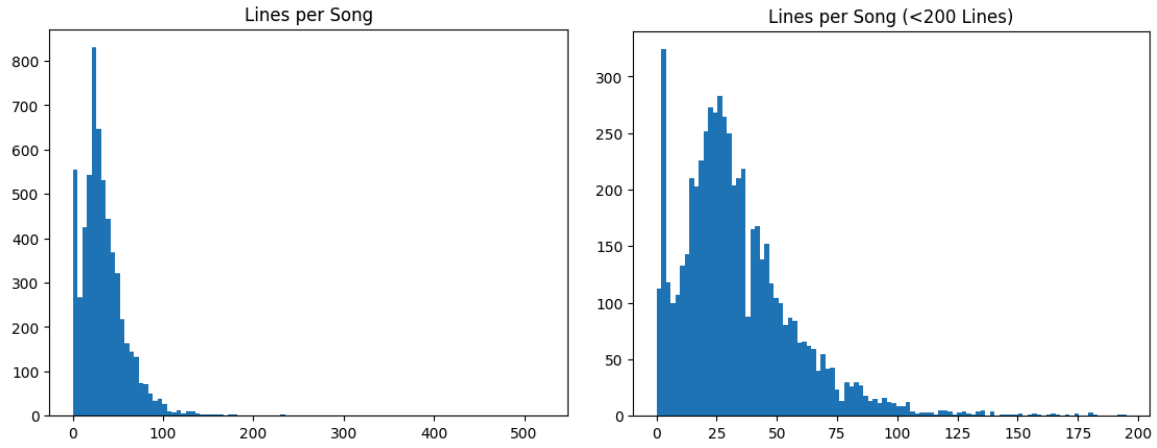
2.2. Data Cleaning

In the preprocessing phase for both tweets and lyrics, each text instance underwent essential cleaning procedures. These procedures involved converting all letters to lowercase, eliminating English stopwords, punctuation, whitespace characters, and numbers. As a final refinement step, lemmatization using the WordNet lemmatizer was applied.

However, it's worth noting a limitation in the song lyrics dataset—specifically, the absence of a breakdown of lyrics into distinct lines. Consequently, we adopted an approach where each consecutive set of 6 words is considered a line of a song. After implementing this breakdown,

an examination of the distribution of each song in terms of line numbers revealed a notable right skewness and instances of songs with a minimal number of lines.

To ensure the integrity of subsequent calculations and prevent skewed results, songs with fewer than 5 or more than 20 lines were removed from the lyrics dataset. This step aims to enhance the reliability and accuracy of the subsequent analyses. For histogram plot of lines per song, please refer to plots below.



3. Modeling

3.1. Tokenization

To develop the best CNN-LSTM model, we initially split the combined tweets dataset into training and testing sets using the train_test_split function from the scikit-learn library. The data is divided with 80% allocated for training and 20% for testing. To maintain a balanced representation of classes, a stratified splitting approach is employed.

For defining the input tensor, a maximum number of words to consider (MAX_NB_WORDS) is set at 15,000. Tokenization is carried out using the Tokenizer class from Keras, transforming texts into sequences of indices representing the most frequent 15,000 words. The tokenizer is fitted on the training data and then applied to both the training and testing datasets.

As sequences resulting from tokenization possess varying lengths, a uniform input for the model is created by padding sequences with zeros until a maximum sequence length (MAX_SEQUENCE_LENGTH) of 20 is reached. This standardization ensures consistency and aids in the processing of sequences during model training and testing.

To prepare the labels for the model, indicating suicidal intention (intention), they are converted into categorical values. The to_categorical function from Keras is utilized for this task, transforming integer labels into a binary matrix representation. This conversion is essential for utilizing categorical cross-entropy, a commonly used loss function for binary classification.

3.2. Model Development

In crafting a model for detecting depressive content in song lyrics, a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture was developed. Implemented using the Keras library, the model architecture includes an embedding layer for converting integer sequences into dense vectors, followed by two convolutional layers with max-pooling and dropout for feature extraction. Afterward, an LSTM layer is employed to capture sequential patterns, with dropout applied to prevent overfitting. The final layer is a dense layer utilizing softmax activation for binary classification into depressive or non-depressive categories.

To optimize the model's performance, a systematic exploration of hyperparameters was conducted through a grid search. Key parameters include optimizer choices (Adam, SGD, RMSprop), dropout rates for the convolutional and LSTM layers (ranging from 0.1 to 0.4), a fixed epoch count of 3, and a batch size of 64.

For each set of defined parameters, the model was trained on the provided training data, and its performance was evaluated using the ROC AUC score on the test data. This score provides insights into the model's proficiency in distinguishing between depressive and non-depressive content within a dataset not encountered during training. The results, encompassing the trained models and their respective ROC AUC scores, are systematically recorded in a dedicated dataframe.

Upon completion of the grid search, the records offer a comprehensive overview of the model configurations, the actual trained models, and their associated ROC AUC scores. A plot depicting the performance of each parameter combination on the test dataset is presented. Notably, the 'sgd' optimizer exhibits inferior performance compared to 'adam' and 'sgd', with both 'adam' and 'sgd' optimizers performing nearly identically across various dropout rate combinations.

The optimal model parameters, as evident from the graph above, are as follows:

- Optimizer: **adam**
- Drop-out Rate for Conv: **0.4**
- Drop-out Rate for LSTM: **0.1**
- Epochs: **3**
- Batch size: **64**
- ROC AUC on Test split: **95.25%**
- ROC AUC in Training Complete Dataset: **98.9%**

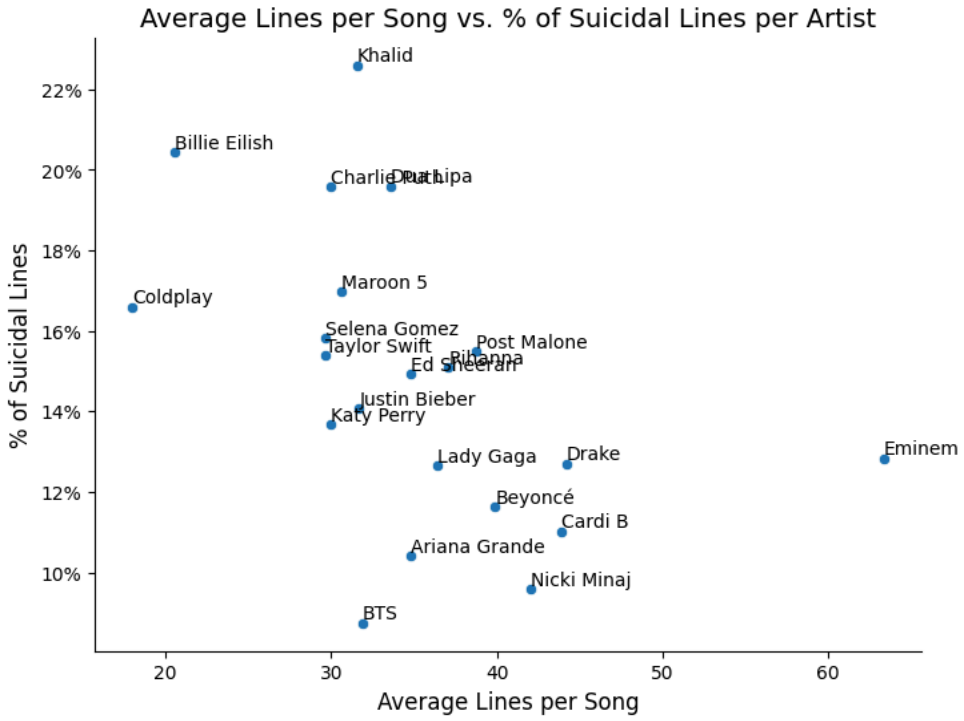
Subsequently, the model is fitted with the entire training dataset to enhance its foundational knowledge. Moving forward, this well-tuned model will be utilized for further analyses, leveraging its robust performance in identifying depressive content in song lyrics.

4. Application on Lyrics

Utilizing the CNN-LSTM model developed earlier, the aim is to unveil patterns and trends that shed light on the presence of depressive or suicidal content within the lyrical compositions of diverse artists. To achieve this objective, the model is applied to label each song present in the lyrics dataset.

4.1. Overall Assessment of Artist' Lyrics

First, a scatter plot for quick and intuitive assessment of the distribution and concentration of potentially concerning lyrical content within the artists' bodies of work is generated. Here, the average line per song represents the number of songs divided by the sum of lines per artist, and the percentage of suicidal lines represents the percentage of lines in the artist's lyrics that is labelled as an expression of depressive or suicidal thoughts.



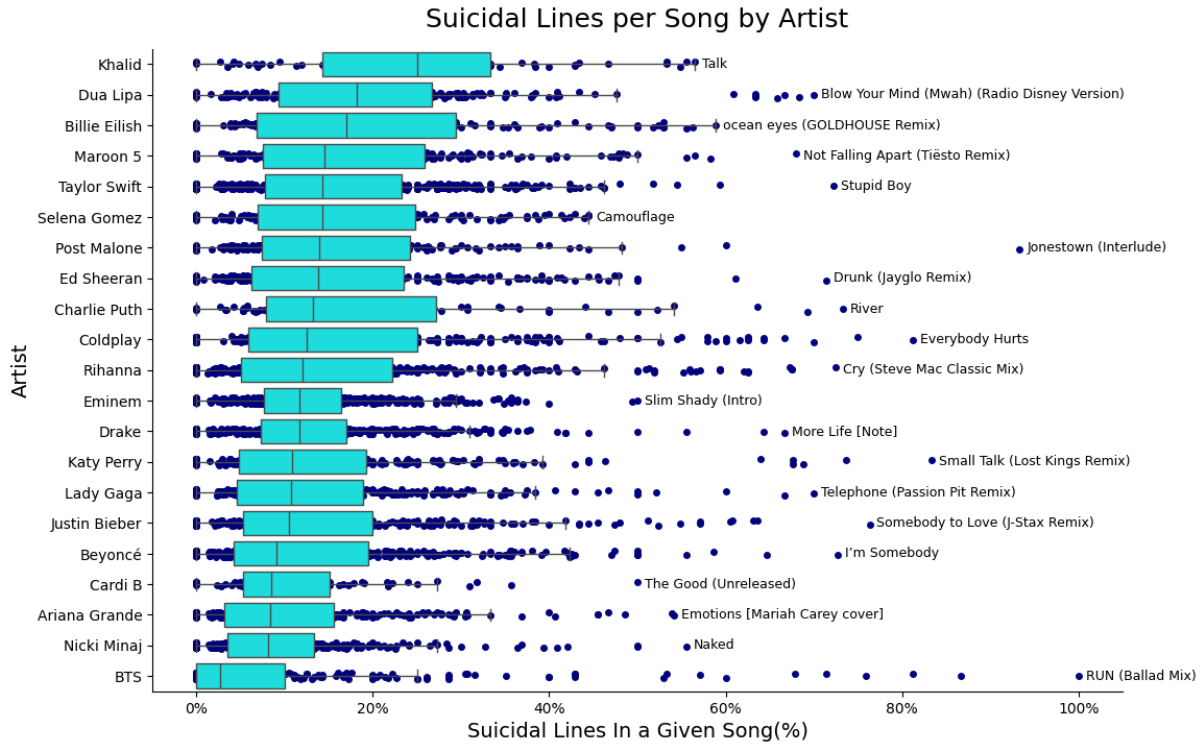
Khalid and Billie Eilish emerge as the artists with the highest percentage of suicidal lines, while BTS and Nicki Minaj exhibit the lowest percentage. Notably, rappers such as Eminem, Drake, Cardi B and Nicki Minaj have the highest number of average lines per song compared to other artists.

Furthermore, a noteworthy observation is a negative correlation between an artist's song length (average lines per song) and the percentage of suicidal lines in their lyrics. This suggests that artists expressing depressive thoughts in their songs tend to have shorter songs overall.

4.2. Assessment on Song Level

In our exploration of song-level data, the primary aim is to comprehensively analyze how artists vary in terms of the prevalence of suicidal content in their song lyrics. The graphical representation is designed to facilitate a comparative analysis, providing insights into the diverse expressions of depression or suicidality within the lyrics of individual artists.

On the y-axis, artists are sorted based on the median percentage of suicidal lines present in their songs. Simultaneously, the x-axis illustrates the percentage of suicidal lines within a given song, with individual dots representing each song by the respective artist. The position of the boxes along the y-axis signifies the median percentage of suicidal line values.



Artists like Khalid, Dua Lipa, and Billie Eilish stand out due to a higher proportion of songs containing suicidal lines. Furthermore, the median percentage of suicidal lines by song for these artists exceeds 15%, indicating that, at least 15% of the lyrics in half of their released songs can be labeled as expressing suicidal intention.

In contrast, the majority of artists exhibit a median percentage ranging between 10% and 15%. An exception is BTS, with a median below 5%, highlighting a more neutral lyrical tone in the majority of their songs.

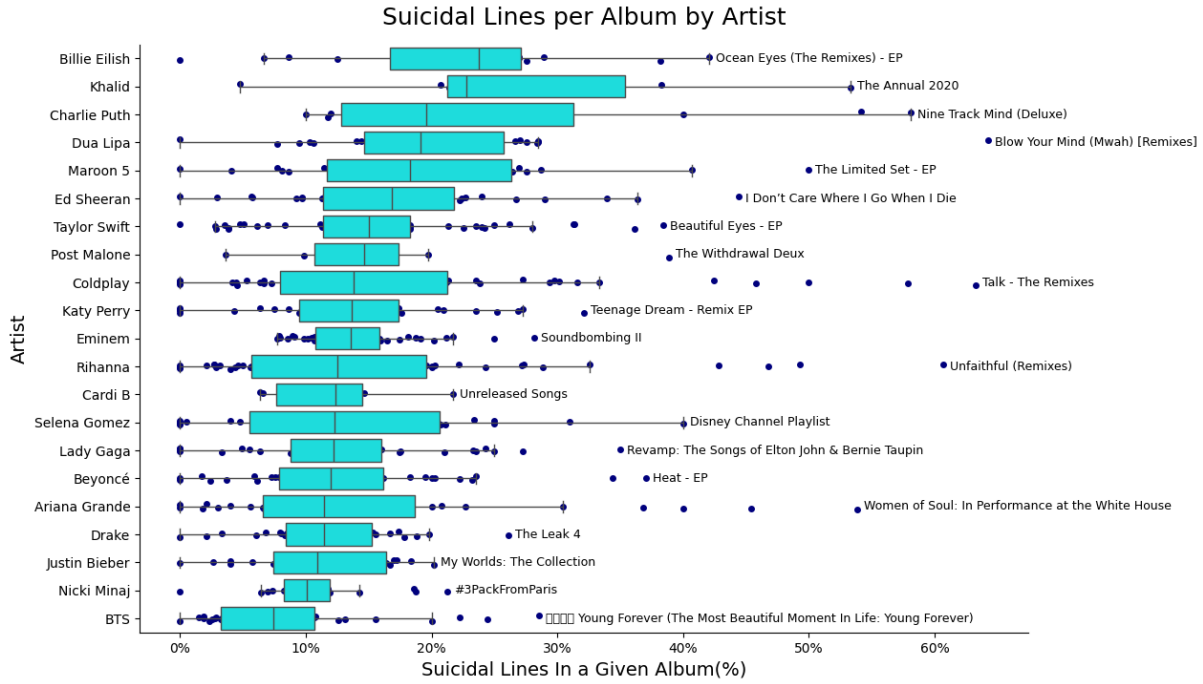
The inclusion of outlier dots allows the identification of songs distinctly labeled as suicidal compared to the general distribution of an artist's songs. Annotations for each artist specify the song title with the highest percentage of suicidal lines. This provides additional insights into specific songs that deviate significantly from the overall pattern of the artist's lyrical content.

4.3. Assessment on Album Level

Extending our analysis to the album level, our primary objective is to develop a comprehensive understanding of how artists vary in terms of the prevalence of suicidal content in their albums. Through the graphical representation, our aim is to facilitate a

comparative analysis, offering insights into the diverse expressions of depression or suicidality across entire albums.

Similar to the song-level analysis, the y-axis sorts different artists based on the median percentage of suicidal lines within all songs in a given album. Conversely, the x-axis illustrates the percentage of suicidal lines within a given album, with individual dots representing each album by the respective artist. The position of the boxes along the y-axis signifies the median percentage values of suicidal lines within albums.



Similar to the breakdown at the song level, artists such as Billie Eilish and Khalid stand out with a larger proportion of albums containing suicidal lines. Remarkably, the median percentage of suicidal lines by album for these artists exceeds 20%, indicating that, at least 20% of the lyrics across half of their albums contain expressions categorized as suicidal intention.

Contrastingly, the majority of artists tend to have median values ranging between 10% and 20%, with the exception of BTS, whose median falls below 10%. This observation suggests that BTS tends to have more neutral lyrics in the majority of their songs across albums.

The presence of outlier dots allows for the identification of albums that are notably labeled as suicidal compared to the general distribution of an artist's albums. Annotations for each artist specify the album title with the highest percentage of suicidal lines. This provides additional

granularity to our analysis at the album level, highlighting specific albums that deviate significantly from the overall pattern of the artist's lyrical content.

5. Conclusion

This project aimed to uncover the complex connections between expressions of depression on social media and the lyrical content of various artists. Our investigation began with the analysis of Twitter data marked with suicidal intention, utilizing a robust Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. The model, refined through meticulous fine-tuning via a grid search, set the foundation for our subsequent analyses. Expanding our focus from social media to song lyrics, we employed this trained model to examine and quantify depressive expressions in the musical compositions of diverse artists, providing a unique insight into how individuals convey and share their emotional struggles.

Our exploration of artists' lyrics revealed compelling results. Notably, artists like Khalid, Billie Eilish, and Dua Lipa had a significant proportion of songs containing suicidal lines. The identification of a negative correlation between an artist's song length and the percentage of suicidal lines added depth to our understanding, suggesting that artists expressing depressive thoughts tend to have shorter songs. Taking our analysis to the album level uncovered nuanced patterns, with Billie Eilish and Khalid continuing to stand out, showcasing a substantial percentage of albums containing suicidal lines, exceeding 20%.

While our approach offers valuable insights, it's essential to acknowledge limitations, such as the subjectivity in labeling depressive content and the constraints of our song lyric breakdown. Future directions may involve refining sentiment analysis models and exploring the impact of music consumption on mental health.

In conclusion, our project resides at the intersection of computational analysis and emotional exploration. Leveraging advanced models, we delved into the world of human expression, shedding light on the prevalence of depressive content in both social media and artists' lyrics.

6. Appendix

For further reference and a detailed overview of the methodologies, code, and data used in this study, please visit the following GitHub repository. This repository contains all the scripts, datasets, and analysis files used in our study. It provides a comprehensive

resource for understanding the text mining techniques applied, including clustering, sentiment analysis, topic modeling, and Word2Vec analysis.

Repository Link: https://github.com/mrcljns/Text-Mining-Projects/tree/master