EMOTION DETECTION IN SONG LYRICS STANZAS

TEXT ANALYTICS - A.Y. 2024/2025

Group 2 - Barbieri, Bosco, Ferrara, Rotellini, Zizza

Abstract

Lyrics serve as one of the main foundations of songs, playing a crucial role in expressing feelings in many different ways. The emotional tone of songs can also serve various purposes, such as automatized playlist creation or songs' organization, offering an alternative to the more traditional genre-based classification. This study focuses on developing three Machine Learning models to classify emotions conveyed in English song lyrics, at the stanza level. The models were chosen for their proven effectiveness across various domains and their diverse approaches, providing a thorough investigation of different techniques and depths for emotion classification in text.

The study begins with the preprocessing of the *Genius Song Lyrics*[1] dataset. The models were then trained through transfer learning, by generating the ground truth using the already existing ALBERT Base v2 model. The results are discussed to highlight the specific challenges encountered.

Introduction

This study focuses on capturing emotional dynamics in song lyrics by assigning emotion labels to individual stanzas rather than entire songs. This approach allows for a more granular analysis of emotional shifts within the text. The emotion labels correspond to Robert Plutchik's eight primary emotions[2], (illustrated in the picture), providing a comprehensive range for representing various emotional states.

The goal of this report is to provide a comprehensive overview of the project, detailing its methodology, results, and key insights. The *Methods* chapter contains a detailed explanation of the data and procedures used in the project, providing descriptions of each part implemented in the project.

The *Results* chapter presents an overview of the obtained outcomes.
These results are further explored in the final sections, *Discussion* and *Conclusions*, which interpret the general findings, present possible future directions and recap the primary objectives of the work.

Methods

The dataset used in this project is a sampled subset of English-language songs derived from the *Genius Song Lyrics Dataset*[1]. The original dataset contained numerous attributes; the ones considered relevant for model training are:

- **title:** the song's title;
- **lemmatized_stanzas:** lyrics of the single stanza;
- **stanza_number:** identifies the position of the stanza in the song;
- **is_chorus:** boolean variable that attests whether the stanza is a chorus or not;
- **is_country, is_pop, is_rap, is_rb, is_rock:** boolean variables, result of a one-hot encoding process, that represent songs genres;
- **label:** represents the emotional classification of the stanza, assigned by Albert Base v2[3].

All of these attributes, except for `title`, were the result of the preprocessing phase, as described in the preprocessing section. Due to limited computational power, the labeling process was time-intensive, ultimately resulting in a limited dataset, with a few more than 100.000 entries.

Preprocessing

The preprocessing phase began by sampling the dataset while maintaining genre distribution. Text cleaning focused on removing irrelevant noise, such as square-bracketed items, and splitting lyrics into individual stanzas using stanza-related keywords (e.g. "chorus", "verse", "bridge" etc.). Resulting uninformative stanzas, such as empty or very short ones, were discarded. The output of this preliminary preprocessing phase was a dataset where the records were no longer whole songs but

individual stanzas, each numbered according to its position within the song. A boolean feature, `is_chorus`, was then added to mark repeated or chorus-related stanzas, and duplicate stanzas were removed to eliminate redundancy. Further cleaning involved removing stanza headers and newline characters, producing cleaner stanzas for labeling.

Lemmatization was performed using the `spaCy` library, generating tokenized stanzas by filtering out punctuation. Lemmatization was chosen over stemming because it produces more accurate and meaningful results, particularly for tasks requiring semantic understanding, such as the one at hand.

Since the dataset was not pre-labeled at the stanza level, ALBERT Base v2 was then fine-tuned to label approximately 100,000 stanzas with emotional categories. A preliminary class distribution analysis showed a slight imbalance, with *joy* being the most frequent (18%) and *disgust* the least (10%). Figure 1 illustrates the distribution across all classes.

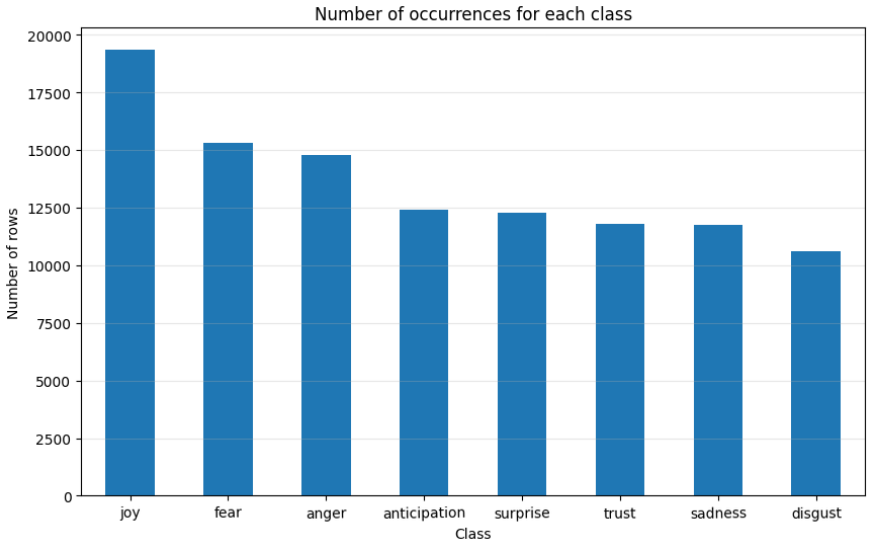


Figure 1: Number of stanzas for each label

Models

The selected model architectures are:

- **Random Forest**: A robust ensemble learning method known for its ability to handle com-

plex, high-dimensional datasets effectively.

- **One-Dimensional Convolutional Neural Network (1D-CNN)**: Designed to capture local patterns in sequential data, leveraging convolutional layers to learn hierarchical features.
- **Recurrent Neural Network (RNN)**: Utilized for its strength in processing sequential data, with the ability to capture contextual relationships between words across different stanzas.

The initial project proposal included a Support Vector Machine as well, but it was discarded as the training was computationally too expensive. The code is still available in the project, both for training and for usage. Their different approaches and depths are an important point of the study, as they offer interesting insights into the possible different techniques and levels of complexity required for detecting emotional tones in complex pieces of text.

Random Forest

The development of Random Forest aimed at providing a benchmark for the more complex neural networks. Its architecture consists of a preprocessing layer, followed by a classifier. The preprocessing handled `title` and `lemmatized_stanzas` using TF-IDF for feature extraction.

To improve classification, feature extraction was conducted on the labeled dataset. Firstly, we deepened the preprocessing by creating a custom stopword list of frequent and generic words, punctuation, and common typographical errors. To this list were then added NLTK's stopwords and numbers lists. Then, TF-IDF scores were computed for each emotion label, using the parameters `min_df` and `max_df` to minimize the influence of overly common or rare words.

Despite these efforts, the analysis did not yield the expected results, with most labels sharing common features and low TF-IDF scores. This was likely due to the repetitive and generic nature of song lyrics.

Random Search was employed for hyperparameter tuning, with 5-fold cross-validation used to provide a more reliable estimate of model performance.

Neural Networks

Neural network architectures were developed and tuned through empirical testing. Both the Recurrent Neural Network and One-Dimensional Convolutional Neural Network share the same preprocessing steps: Non-Negative Matrix Factorization is applied to the title for extracting latent topics, following TF-IDF for richer text representation. The `lemmatized_stanzas` are processed through convolutional and recurrent pipelines, where elements are tokenized and padded to maintain consistent input shapes.

One-Dimensional Convolutional Neural Network

The Convolutional part of the architecture is specifically designed to extract and learn local patterns in `embedding_lyrics`. Its structure consists of three convolutional layers, each applying filters of varying sizes. This allows to detect patterns at different granularities. These layers are followed by Global Max Pooling, to reduce the previous output's dimension to a fixed-length vector, as well as retaining focus on the most informative patterns. A dropout layer is then applied, to introduce regularization and prevent overfitting.

Recurrent Neural Network

The Recurrent part of the architecture is specifically designed to extract and learn local patterns in `embedding_lyrics`. Its structure consists of three Gated Recurrent Units (`GRU` layers) to model temporal relationships. These are characterized by progressively smaller numbers of units; this allows pattern capture at different abstraction levels. All three layers in the architecture use the `tanh` activation function to compute the hidden state and the `sigmoid` activation function for the recurrent gate. The first and second layers return the full sequence of hidden states for each time step in the input sequence, enabling richer learning of patterns over time. Dropout is applied on every layer, to prevent overfitting and add regularization.

Shared Components

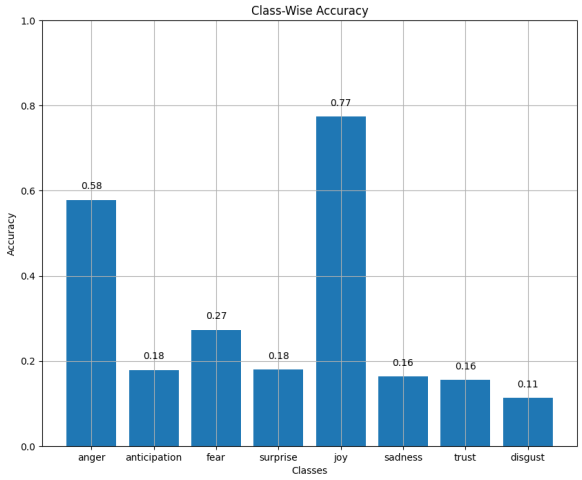
Other features are processed through a simple pipeline, which concatenates inputs, passes them through a dense layer, and combines them with the lyrics-processing output. The final output layer

uses 8 units with softmax activation for emotion classification.
The models are trained using categorical cross-entropy as the loss function, and categorical accuracy as the evaluation metric. Other metrics, such as top-k categorical accuracy with $k = 2$, were tested but discarded due to limited performance improvements and lower precision.

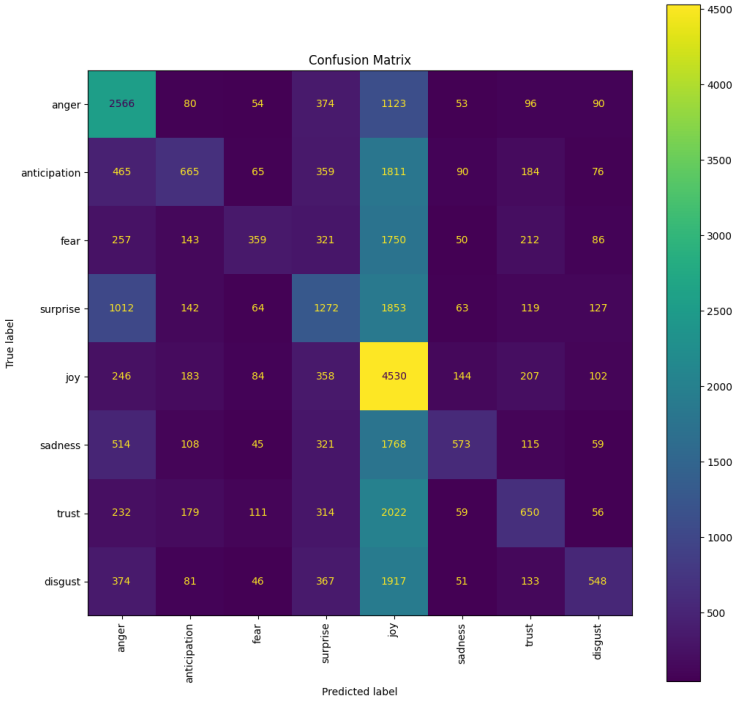
Results

The performance metrics considered for evaluation in these analysis are the following: accuracy, precision, recall, and F1-score. Considering all of these metrics is crucial to accurately evaluate how well each model performs. The classification report revealed an accuracy of 34% for Random Forest. This result can be considered reasonable, given the fact that the task at hand is a multi-class classification problem with 8 classes.

For the Random Forest model, the class with the highest F1-score is *anger*, at 51%, which is the third class as for support, sitting at 4436. *Joy* was the most supported class, with 5854 instances, but it came second as for the F1-score, which was of 0.40. The second class based off support, which was *fear* with 4652 instances, had a lower F1-score of 0.31. The remaining classes showed comparable support and F1-scores, averaging around 3500 instances and 0.25 respectively, only *disgust* had a considerably lower F1-score, at 18%.
In image **??**, the confusion matrix of Random Forest is displayed. It is clear how the best classified classes are *joy* and *anger*. However, the most interesting insight that can be gathered from the confusion matrix is the extent to which all other classes get misclassified as *joy*. This is probably because *joy* is the dominant class, due to an imbalanced dataset. Furthermore, as was highlighted in previous sections, the most important features for classification are significantly shared between the classes, therefore making it harder for the model to classify correctly. Then, image 2a shows class-wise accuracy.



(a) Class-wise Accuracy - Random Forest

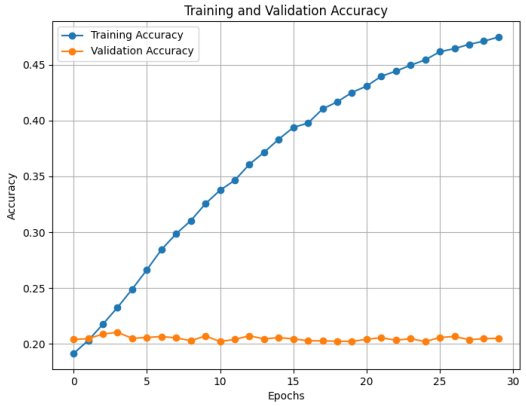


(b) Confusion Matrix for the Random Forest Classifier

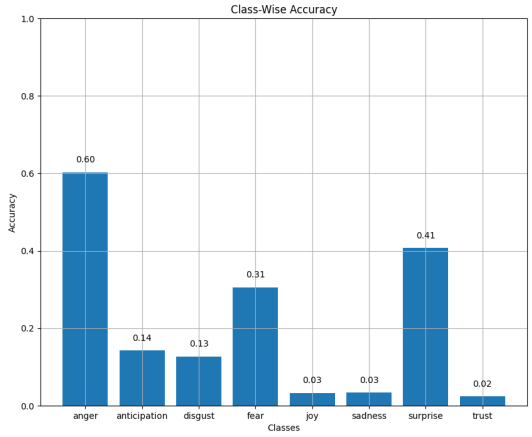
Figure 2: Random Forest - plots

As mentioned in the previous chapter, the development of neural networks iterated testing phases and adjustments over different aspects of training. Semi-supervised learning through generation of pseudo labels via the partially trained models generally yielded poor results; on the other hand, downsampling into evenly represented labels gave some minor improvements, for both architectures. The neural networks generally underperformed compared to Random Forest: the convolutional neural network ultimately reached a test categorical accuracy of 0.2128, while the recurrent neural network had a test accuracy of 0.1869.

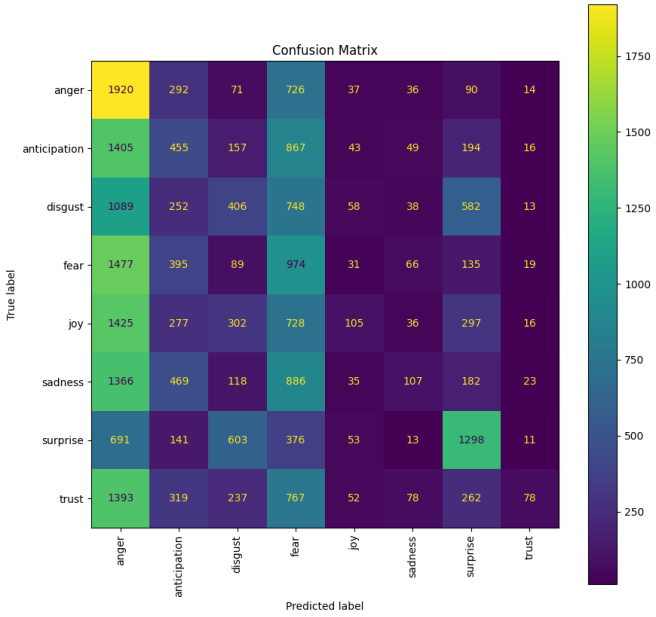
The graphs below show various performance metrics of the network with the better performances for the convolutional architecture.



(a) Training and validation accuracy



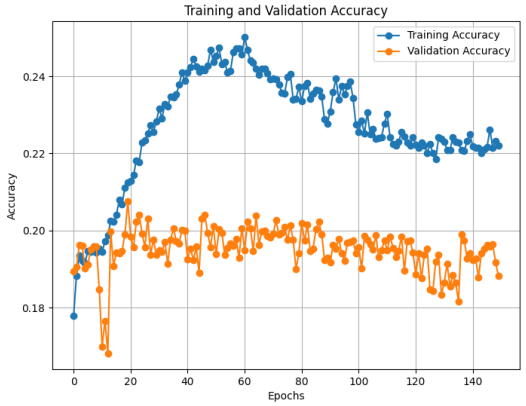
(b) Class-wise accuracy



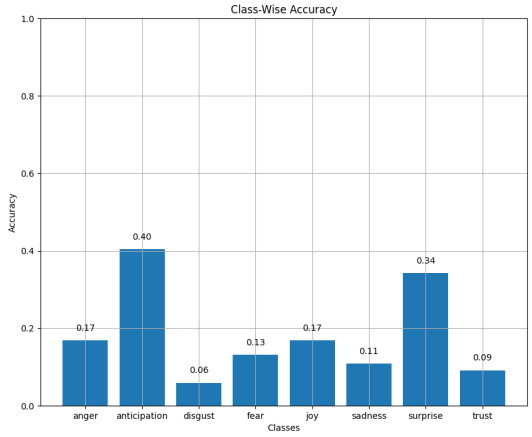
(c) Confusion matrix

Figure 3: Convolutional Neural Network - plots

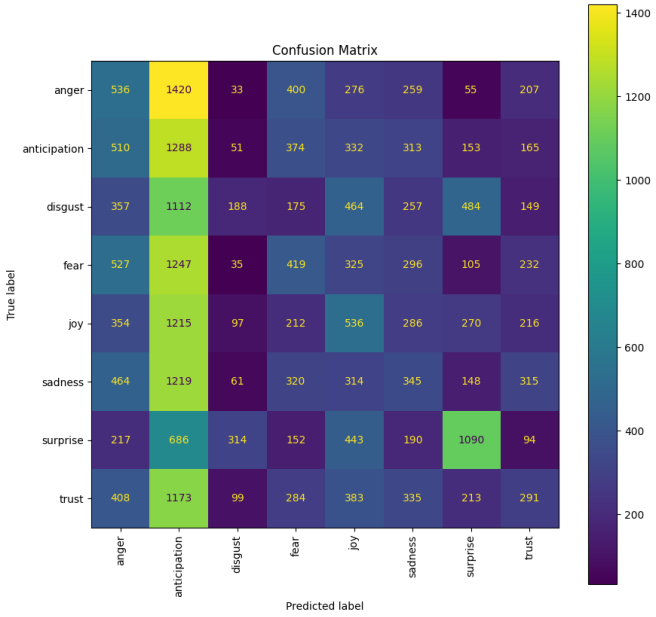
The graphs below show various performance metrics of the network with the better performances for the recurrent architecture.



(a) Training and validation accuracy



(b) Class-wise accuracy



(c) Confusion matrix

Figure 4: Recurrent Neural Network - plots

Discussion

The results presented in the previous chapter show general uncertainty in the models. This has a few possible explanations: one of the most likely is the significant overlap of features. This might explain why all class-wide accuracy plots show the predominance of one or two classes over the rest. In different trainings, the resulting predominant classes seem to be chosen at random.

There are clear indications of this issue in the recurrent network's performances. The model presented in the previous chapter shows a general low training accuracy (figure 4a), but this was the configuration that yielded better class-wise accuracies (figures 4b, 4c). With a lower number of epochs, the two predominant classes tendency is worsened, as shown below for an 80 epochs model (figure 5).

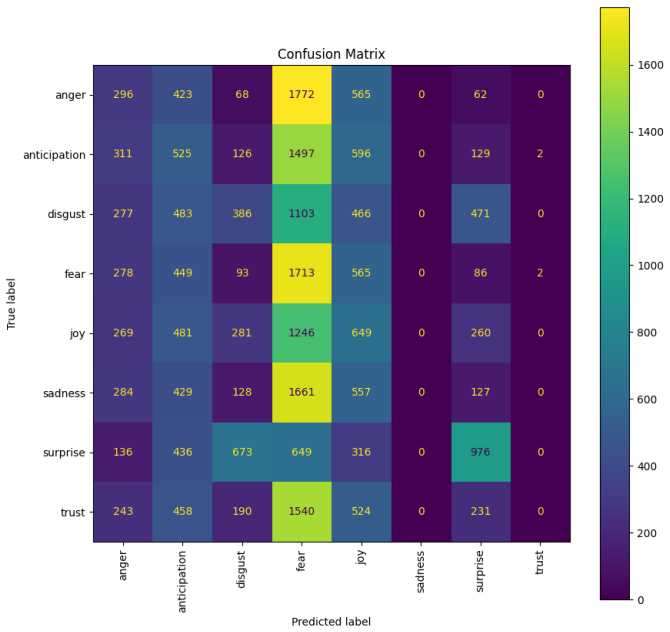


Figure 5: Confusion matrix for a 80-epochs recurrent neural network

It is possible that with higher numbers of epochs the model continues improving on lower performing classes, but for computational reasons this was not tested.

The convolutional architecture reaches a better training accuracy, but with minimal corresponding gains in validation accuracy (figures 3a). To a lesser extent, the model also shows the two predominant classes tendency, as seen in the confusion matrix and class-wise accuracy plot (figures 3b, 3c).

Random Forest has better perfomances in this metric as well, although the issue is still present (as seen in figures 2a, 2b).

To address this issue, one solution could be applying a filter to exclude the most common words, allowing the remaining words to better convey clear emotional meanings. Another potential unexplored approach is the use of custom metrics during training that focus on class-wise accuracy instead of overall model accuracy.

Another probable issue is a flawed ground truth, which may also contribute to the poor performance and could be linked to the vague distinction between classes. This possibility can be investigated through explainability analysis, with an example provided below (figure 6).

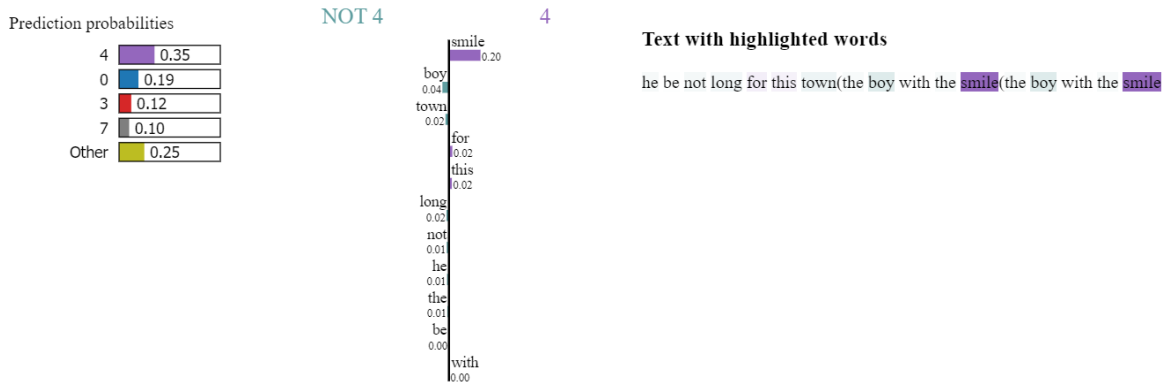


Figure 6: Explainability - visualization

The left section of the graph displays the predicted probabilities for each class. In the center section, feature importances are ranked from most to least relevant and divided into two groups: on

the right, features with a positive influence on the predicted label; on the left, those with a negative influence that suggest the model should consider other classes. The right section highlights the values of the most important features, using bright colors to indicate features with a positive influence on the prediction. The example in question illustrates a prediction where the model assigned the label *joy* to the stanza under analysis, but the expected label, assigned by the ALBERT model, was *sadness*. However, the word *smile*, which is brightly highlighted, intuitively suggests that *joy* might be a more plausible class for this stanza, even one that ALBERT could reasonably assign. This showcases the aforementioned issue: the transfer learning approach used to create the ground truth appears to have some limitations. In some instances, labels generated by ALBERT do not seem to be appropriate.

One possible solution is to switch to a different pseudo-labeling model. Alternatively, using probability vectors rather than direct classification for labeling could guide the models toward a regression-based approach; this also allows the ground truth-generator model to provide more context, providing potentially useful inputs for the models and facilitate a deeper understanding of their performance issues, helping to identify where and why they struggle.

Conclusions

This study aimed at exploring various Machine Learning techniques perform an emotion detection task on songs, which are irregular, complex texts. The particular field has many practical applications, such as improving recommendation systems.
The results of the project point towards better performances obtained by the more straightforward, simpler model; neural networks generally struggled to find general, meaningful patterns and correlations, leading into suboptimal training and testing performances. These results might have been caused by a series of factors, such as the likely feature overlap between different classes and unreliable labeling.

As mentioned in the previous chapter, there are things that can be done to solve both these issues, such as using alternative models, techniques or sources for generating the ground truth. Another

possible approach is to use a probabilistic approach for the labeling process, which can indeed guide models into more informed decisions.

In conclusion, emotion detection can improve by refining techniques and enhancing the quality of data. Addressing key issues such as feature overlap and unreliable labels could lead to more robust models, ultimately improving their usefulness in real-world applications such as music recommendation systems and content curation across various media platforms.

Bibliography

[1] *Genius Song Lyrics*. URL: https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information?select=song_lyrics.csv.

[2] Robert Plutchik. "A general psychoevolutionary theory of emotion". In: *Emotion: Theory, research, and experience* 1 (1980).

[3] *Albert Base v2*. URL: https://huggingface.co/albert/albert-base-v2.

List of figures