Music Emotion Recognition (MER): Valence Prediction

This project focuses on **Music Emotion Recognition (MER)**, aiming to determine whether the emotional content of songs can be automatically detected. In particular, we address the **prediction of valence**, a continuous emotional dimension representing how happy or sad a song is. Accurately modeling valence is especially relevant for music recommender systems, as it enables recommendations that take into account not only users' listening history but also their desired emotional state.

The task is formulated as a **regression problem**, and we evaluate multiple modeling strategies with increasing levels of complexity. These approaches differ both in the type of textual representation used and in the learning architecture applied, allowing for a comprehensive comparison between traditional NLP methods and modern transformer-based models.

Traditional NLP Embeddings with Simple Regressors

As a first baseline, we use traditional NLP embeddings extracted from song lyrics, including TF-IDF, n-grams, Word2Vec, Doc2Vec, GloVe, FastText, and LDA-based topic representations. These embeddings are combined with **simple regression models**, primarily Ridge regression.

Results (Pearson Correlation – Ridge)

Approach	Lyrics Only	Lyrics + Audio
TF-IDF	0.291	0.594
TF-IDF + n-grams	0.307	0.601
Word2Vec	0.218	0.608
Doc2Vec	0.251	0.613

GloVe	0.279	0.613
FastText	0.306	**0.623**
LDA (various topics)	~0.21–0.22	~0.61

These results show that lyrics alone contain meaningful emotional information (Pearson ≈ 0.20–0.31). However, a clear and consistent improvement is observed when audio features are added, increasing performance to approximately 0.60–0.62 across most approaches. This demonstrates that acoustic information provides complementary emotional cues that are not fully captured by lyrics alone.

For these simpler models, text preprocessing plays an important role. Removing repetitive interjections such as *"yeah"* or *"oh"* generally helps performance, as these expressions tend to be interpreted as noise rather than informative content.

Traditional NLP Embeddings with Advanced Regressors

To isolate the effect of model capacity, we next combine traditional NLP embeddings with **advanced regression architectures**, such as multilayer perceptrons (MLPs), while keeping the textual representations fixed.

Results (Pearson Correlation)

Model	Pearson
FastText + Audio + MLP (`ft-emb_audio_mlp`)	**0.579**

Using a more expressive regressor leads to a noticeable performance improvement compared to simple linear models applied to the same embeddings. Nevertheless, the results remain below those achieved by end-to-end transformer-based models, suggesting that the main limitation of these approaches lies in the fixed nature of the embeddings rather than in the regression architecture itself.

Transformer Embeddings with Simple Regressors

In the next stage, we evaluate **transformer-based embeddings** extracted from pre-trained language models such as **BERT, RoBERTa, and XLNet**. These embeddings are treated as fixed representations and combined with simple regression models.

Results (Pearson Correlation – Ridge)

Approach	Lyrics Only	Lyrics + Audio
BERT embeddings	0.306	0.612
RoBERTa embeddings	0.305	**0.614**
XLNet embeddings	0.293	0.612
Clean BERT	0.271	0.608

Transformer embeddings outperform most traditional text-only baselines, but the gains are moderate when they are used only as static feature extractors. Once again, adding audio features yields a large and consistent improvement, pushing performance above 0.61 Pearson.

Interestingly, cleaning the lyrics does not improve results in this setting. This suggests that even when used as fixed embeddings, transformers are relatively robust to informal expressions and repetitive patterns.

Full Transformer Models (End-to-End)

Finally, we evaluate **full transformer models trained end-to-end**, where the transformer directly processes the lyrics and is optimized jointly with the valence prediction objective.

Results (Pearson Correlation)

Lyrics Only

Model	Pearson
raw-BERT	**0.410**
raw-RoBERTa	0.399
raw-XLNet	0.362

Lyrics + Audio Features

Model	Pearson
raw-BERT + audio	0.636
raw-RoBERTa + audio	0.635
raw-XLNet + audio	**0.640**
FastText + audio + MLP	0.579

Full transformer models significantly outperform all previous approaches. Even with lyrics only, they achieve substantially higher performance than any model based on fixed embeddings. When audio features are included, they reach the highest performance overall, with Pearson correlations close to 0.64.

A key observation is that transformer-based models perform best when trained on raw lyrics, without removing informal expressions such as *"yeah"* or *"oh"*. Due to their higher representational capacity and ability to model long-range dependencies, transformers can interpret these elements as stylistic or emotional signals rather than noise. In contrast, simpler models tend to benefit from cleaner text, as they lack the capacity to distinguish meaningful repetition from irrelevant noise.

Conclusion

The experimental results reveal a clear progression in performance as model complexity increases. Traditional NLP embeddings with simple regressors provide strong and interpretable baselines, while advanced regressors yield moderate improvements. Transformer embeddings further enhance performance, but the largest gains are achieved by full transformer models trained end-to-end, particularly in multimodal settings.

Across all approaches, the inclusion of audio features consistently improves performance, confirming the importance of jointly modeling textual and acoustic information for valence prediction. Furthermore, the results highlight that preprocessing strategies should be model-dependent: simpler models benefit from cleaner lyrics, while transformer-based architectures achieve better results when trained on raw textual data.

Overall, these findings demonstrate the effectiveness of **multimodal, end-to-end transformer models** for **Music Emotion Recognition**, and point to promising directions for emotion-aware music recommender systems.