TRANSFORMER-BASED APPROACH TOWARDS MUSIC EMOTION RECOGNITION FROM LYRICS

Authors – Yudhik Agrawal , Ramaguru Gurur Ravi Shanker and Vinoo Alluri

Review on the research paper

Yomal De Mel

Introduction

- Music Emotion Recognition (MER): Importance in Music Information Retrieval (MIR)
- Traditional Methods: Acoustic features, social tags
- Focus: Role of lyrics in MER using transformer-based models

Background

- Previous Studies: Limited use of lyrics in MER
- NLP Techniques: Traditional vs. modern (transformers)
- Objective: Use XLNet for emotion recognition from lyrics

Datasets

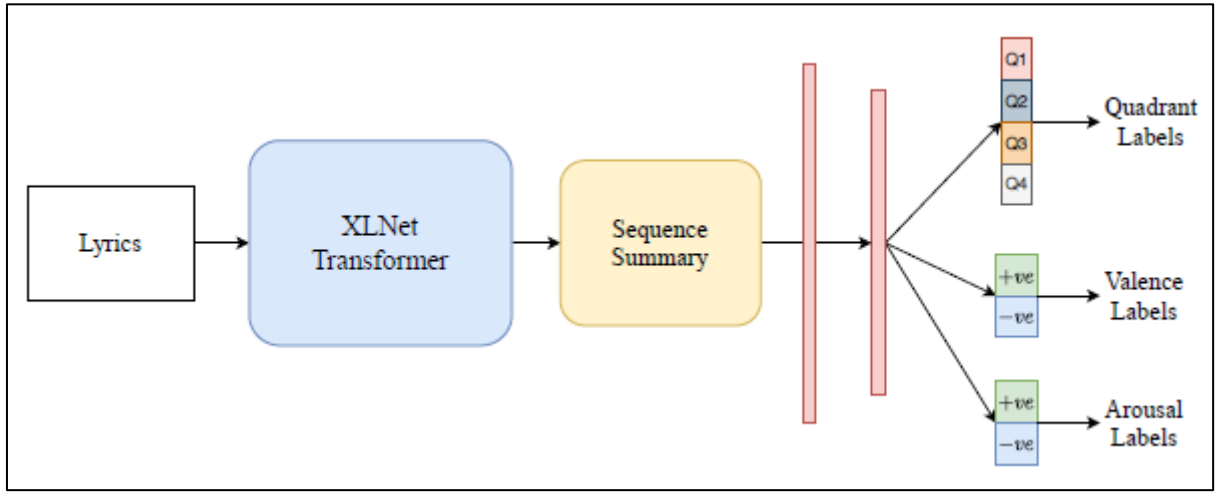
- **MoodyLyrics**: 2595 songs, Valence-Arousal (V-A) model
- **MER Dataset**: 180 songs, annotated V-A values
- **Lyrics Extraction**: Robust methodology using Genius API

Methodology

- **Model Architecture**: XLNet as base, multi-task learning
- **Tasks**: Classification of emotion quadrants, valence, and arousal
- **Training**: AdamW optimizer, cross-entropy loss, dropout regularization

Proposed Architecture

- **Overview**: XLNet transformer model
- **Sequence Summary Block**: Computes vector summary
- **Classification Layers**: Quadrant, valence, arousal



Evaluation Measures

- **Metrics**: Precision, recall, F1-score
- **Macro-averaged F1**: Robust towards error type distribution

$$F1_x = 2\frac{P_x R_x}{P_x + R_x}; \qquad \mathcal{F}_1 = \frac{1}{n}\sum_x F1_x$$

Results on MoodyLyrics

- Comparison: Naive Bayes, BiLSTM + Glove, Proposed Method
- Performance: Accuracy, precision, recall, F1-score

Approach	Accuracy	Precision	Recall	$\mathcal{F}_1$-score
Naive Bayes [2]	83.00%	87.00%	81.00%	82.00%
BiLSTM + Glove [2]	91.00%	92.00%	90.00%	91.00%
Our Method	94.78%	94.77%	94.75%	94.77%

Results on MER Dataset

- Comparison: CBF + POS tags, Proposed Method
- Performance: Accuracy, precision, recall, F1-score

Classification	Approach	Accuracy	Precision	Recall	$\mathcal{F}_1$-score
Quadrant	CBF + POS tags, Structural and Semantic features [24]	-	-	-	80.10%
Quadrant	**Our Method**	88.89%	90.83%	88.75%	88.60%
Valence	CBF + POS tags, Structural and Semantic features [24]	-	-	-	90.00%
Valence	**Our Method**	94.44%	92.86%	95.83%	93.98%
Arousal	CBF + POS tags, Structural and Semantic features [24]	-	-	-	88.30%
Arousal	**Our Method**	88.89%	90.00%	90.00%	88.89%

Ablation Study

- Single-task vs. Multi-task: Performance comparison
- Training Speed: Faster convergence with multi-task setup

Classification	Accuracy		$\mathcal{F}_1$-score	
	Multi-Task	Single-Task	Multi-Task	Single-Task
Quadrant	94.78%	95.68%	94.77%	95.60%
Valence	95.73%	96.51%	95.67%	96.46%
Arousal	94.38%	94.38%	94.23%	94.35%

Discussion and Conclusion

- Superior Performance of Transformer-Based Approach
- Multi-Task Learning Advantage
- Robust Lyrics Extraction Methodology
- Implications for Music Recommendation Systems