From Text to Lyrics: Evaluating the Performance of Sentiment Analysis Models on Musical Content

NGUYEN Dang Hoang Khang
PAVLOVA Kseniia, YING Beining

University Paris-Saclay

February 14, 2025

Table of Contents

1 Problem
- Introduction
- Datasets

2 Method
- Data Processing
- Models

3 Conclusion

Table of Contents

1 Problem
Introduction
Datasets

2 Method
Data Processing
Models

3 Conclusion

Introduction

Objective: Explore the effectiveness of NLP models trained on generalized datasets in addressing domain-specific tasks.

Approach:

- Fine-tune a sentiment classification model trained on **Reddit comments** (GoEmotions).
- Map its predictions to the **Geneva Emotion Music Scales (GEMS-9)**.

Validation:

- Evaluate the model's performance on the **Lyrics Emotion Dataset**.
- Assess its applicability to **music-related sentiment analysis**.

GoEmotions Dataset

Source: 58K carefully curated **Reddit comments**

Labels: 27 fine-grained emotions + Neutral

Annotation: Human-labeled

Dataset Split:

- **Training:** 43,410
- **Testing:** 5,427

Emotion Categories: Admiration, Amusement, Anger, Annoyance, Approval, Caring, Confusion, Curiosity, Desire, Disappointment, Disapproval, Disgust, Embarrassment, Excitement, Fear, Gratitude, Grief, Joy, Love, Nervousness, Optimism, Pride, Realization, Relief, Remorse, Sadness, Surprise

Lyrics Emotion Dataset

Source: 1,160 song lyrics, manually annotated

Purpose:

- Evaluate whether an **NLP model trained on general text (Reddit)** can generalize to **lyrics-based emotion classification**.

Emotion Categories (GEMS-9): Amazement, Calmness, Joyful Activation, Nostalgia, Power, Sadness, Solemnity, Tenderness, Tension

Key Differences

Comparison of GoEmotions and Lyrics Emotion Dataset

Feature	**GoEmotions**	**Lyrics Emotion**
Domain	Reddit comments	Song lyrics
Granularity	Fine-grained (27 categories)	High-level (9 categories)
Dataset Size	58K (Train 43,410, Test 5,427)	1,160 song lyrics
Purpose	General NLP sentiment classification	Evaluating general models on lyrics emotion recognition

université PARIS-SACLAY

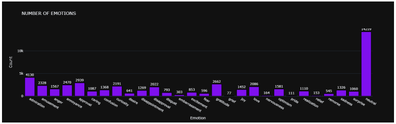
Table of Contents

1 Problem
Introduction
Datasets

2 Method
Data Processing
Models

3 Conclusion

Exploratory Data Analysis: GoEmotions

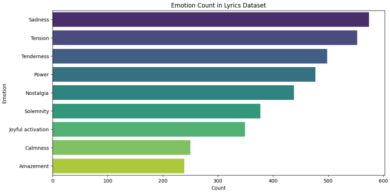


Distribution in GoEmotions

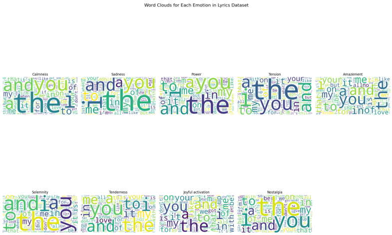
Word Cloud of GoEmotions

- The emotion distribution indicates a significant presence of **neutral** and **approval-related** emotions.
- The word cloud highlights key emotion-specific words, offering insights into sentiment categorization.

Exploratory Data Analysis: Lyrics Emotion Dataset



Distribution in Lyrics Dataset



Word Cloud of Lyrics

- The emotion distribution in the lyrics dataset reveals that **sadness** and **tension** dominate.
- The word cloud highlights frequent lyrical terms associated with different emotional categories.

Data Preprocessing

Steps in Data Preprocessing:

- **Text Cleaning:** Removing special characters, lowercasing, and stripping whitespaces.
- **Tokenization:** Splitting text into individual words or subwords.
- **Stopword Removal:** Eliminating common but uninformative.
- **Stemming:** Reducing words to their base or root form.
- **Label Mapping:** Converting categorical labels into numerical.
- **Train-Test Split:** Ensuring a balanced dataset split for evaluation.

Preprocessing ensures the text data is structured, clean, and ready for model training.

Model Selection

Baseline Model:

- TF-IDF tokenization + Support Vector Machine (SVM) (linear kernel)

Advanced Model:

- RoBERTa (Robustly Optimized BERT Approach) for classification

We compare a simple traditional machine learning model (SVM) with a deep learning approach (RoBERTa) to evaluate their effectiveness on emotion classification.

Baseline Model: TF-IDF + SVM

Method:

- Convert text data into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency).
- Train a **linear SVM** classifier on these feature representations.

Performance on GoEmotions Test Set:

Metric	Precision	Recall	F1-score
Macro Avg	0.68	0.25	0.32
Weighted Avg	0.72	0.35	0.42
Samples Avg	0.39	0.37	0.38

The SVM model shows moderate precision but suffers from low recall, indicating difficulty in capturing minority emotion classes.

Advanced Model: RoBERTa for Emotion Classification

Why RoBERTa?

- Uses **Byte-Pair Encoding (BPE)** tokenization, allowing it to generalize well to unseen words (OOV).
- Better handling of multi-language text.
- Stronger context-awareness compared to traditional models.

Performance on GoEmotions Test Set:

Metric	Precision	Recall	F1-score
Macro Avg	0.57	0.39	0.45
Weighted Avg	0.66	0.51	0.56
Samples Avg	0.57	0.54	0.55

RoBERTa outperforms SVM in recall and F1-score, showing stronger generalization for emotion classification.

Comparison: RoBERTa vs. Baseline (SVM)

Key Observations:

- **RoBERTa** achieves **higher recall and F1-score**, improving the model's ability to detect emotions.
- **SVM** has a higher precision but struggles with recall, leading to poor performance in minority emotion classes.
- **Deep learning (RoBERTa)** captures contextual meaning better than traditional TF-IDF-based approaches.

Performance Summary:

Model	Precision	Recall	F1-score
SVM (Baseline)	0.68	0.25	0.32
RoBERTa	0.57	0.39	0.45

Mapping 27 Emotions to 9 High-Level Categories

Motivation:

- GoEmotions dataset contains **27 fine-grained emotions**, while the Lyrics Emotion Dataset uses **9 broader emotion categories (GEMS-9)**.
- We need to map detailed emotions to higher-level categories to make them comparable.

Emotion Mapping Strategy:

GEMS-9 Emotion	Mapped GoEmotions Categories
Nostalgia	admiration, sadness, love, realization
Tension	fear, nervousness, annoyance, anger
...	...

Mapping ensures compatibility between datasets, enabling cross-domain sentiment analysis.

SVM Performance on Lyrics Emotion Dataset

Key Performance Metrics:

- **Jaccard Similarity Score:** 0.2935
- **F1-Score (Micro):** 0.4215
- **F1-Score (Macro):** 0.2895

Insights from Classification Report:

- **Macro F1-score is low (0.2895)**, suggesting poor performance in capturing minority classes.
- **Micro F1-score is slightly better (0.4215)**, but still suboptimal for a real-world application.

SVM fails to capture complex relationships between lyrics and emotions, likely due to TF-IDF's inability to handle contextual meanings.

RoBERTa Performance on Lyrics Emotion Dataset

Key Performance Metrics:

- **Jaccard Similarity Score:** 0.3509
- **F1-Score (Micro):** 0.4829
- **F1-Score (Macro):** 0.3767

Insights from Classification Report:

- **Significant improvement in Macro F1-score (0.3767 vs. 0.2895)**, indicating that RoBERTa handles diverse emotions better.
- **Lower Hamming Loss (0.3283 vs. 0.3820)**, meaning fewer incorrect labels assigned per sample.

RoBERTa outperforms SVM in recognizing multi-label emotions, thanks to its contextual understanding and subword tokenization.

Performance Comparison: SVM vs. RoBERTa

Key Observations:

- **RoBERTa achieves higher recall and F1-score**, demonstrating its ability to capture nuanced emotions in lyrics.
- **SVM struggles with multi-label classification**, as shown by its high Hamming Loss and low Macro F1-score.
- **Jaccard Similarity and Micro F1-score improvements** indicate that RoBERTa assigns more relevant emotion labels.

Performance Summary:

Metric	SVM	RoBERTa
Jaccard Similarity Score	0.2935	0.3509
F1-Score (Micro)	0.4215	0.4829
F1-Score (Macro)	0.2895	0.3767

Table of Contents

1 Problem
  Introduction
  Datasets

2 Method
  Data Processing
  Models

3 Conclusion

Evaluating Our Approach

Objective & Approach:

- Fine-tune a sentiment model trained on **Reddit (GoEmotions)**.
- Map predictions to **GEMS-9** and validate on **Lyrics Emotion Dataset**.
- Assess applicability to **music-related sentiment analysis**.

Findings:

- **General NLP models maybe struggle** with domain-specific tasks without fine-tuning.
- Even deep models show **suboptimal accuracy** in cross-domain sentiment classification.
- **RoBERTa outperforms SVM** on lyrics, demonstrating better generalization.

Team Contributions

Research and Dataset Selection: Khang, Kseniia, Beining

Data Exploration & Preprocessing: Kseniia

Baseline Model (TF-IDF + SVM): Kseniia, Khang

Advanced Model (RoBERTa): Khang, Beining

Report: Khang (Lead), Beining, Kseniia

Slide: Beining (Lead), Khang, Kseniia

Key Takeaways

What We Learned:

- Hands-on experience with **NLP models for sentiment**.
- **Understanding the challenges** of applying a generalized model to a specific domain.
- The importance of **fine-tuning deep learning models** for better adaptation to domain-specific tasks.
- The limitations of **traditional methods** (like TF-IDF + SVM) when handling nuanced sentiment in complex datasets.
- The role of **contextual embeddings** (like RoBERTa) in improving multi-label classification tasks.

Beyond the technical aspects, we also gained a broader perspective on the strengths and weaknesses of current NLP methodologies in real-world applications.

Future Work

Potential Directions for Improvement:

- **Fine-tuning RoBERTa further** using domain-specific datasets (e.g., emotion-labeled lyrics corpora).
- **Hybrid Models:** Combining traditional methods (TF-IDF, SVM) with deep learning for improved interpretability and performance.
- **Exploring alternative embeddings:** Investigating models like GPT, BERT variants, or multimodal approaches incorporating musical features.
- **Expanding the dataset:** Including more diverse lyric samples and other sources of emotional text.
- **Understanding annotation bias:** Evaluating how human annotations influence model predictions.