# Detecting Emotions in Song Lyrics

Anam Khan (ak1963), Harrison Lee (hpl11), Swaminathan Venkateswaran (sv526)

Prof. Merkhofer– ANLY521 – Computational Ling Adv Python

## Objective

Song lyrics can contain a multitude of emotions (sadness, nostalgia, etc.) and can powerfully portray the emotional state of an artist. This project compares two feature sets to assess whether **content** or **style** is more suitable as a feature set in **multi-emotion classification for song lyrics**.

## Hypothesis

We hypothesize that in the task of multi-emotion classification from song lyric data, content-based features are more important in determining emotions accurately as opposed to style-based features.

## Dataset

This project uses a dataset containing 1,160 songs collected from Genius spanning four of the most popular genres at the time of data collection - rock, country, rap/hip-hop, and reggae. Emotions were hand annotated in 9 emotional classes using the geneva emotional music scale (GEMS). Annotation guidelines can be found under Reference 3

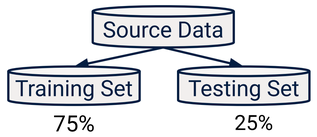


**Figure 1** – An overview of our training and testing set.

## Background

Current literature around the multi-emotion classification problem[1] for song lyrics compares classification algorithms, different emotion labels, as well as textual vs. audio analysis. This project uses the concept of comparing topic or style feature sets[2] for application in the multi-emotion classification problem.

Following prior research, Logistic Regression and Naive Bayes classifiers are used as models for comparing the predictive performance of the two feature sets on classifying the following emotions:

| | | |
|---|---|---|
| **Amazement** | **Calmness** | **Joyful** |
| **Nostalgia** | **Power** | **Sadness** |
| **Solemnity** | **Tenderness** | **Tension** |

## Methodology

### Data Preprocessing:

- Unnecessary columns are removed
- Using CountVectorizer, the single “emotions” column is transformed from a string into a binary representation across 9 columns.
  - (“sadness, nostalgia, tenderness”) → [0, 0, 1, 0, 0, 1, 0, 1, 0]

### Feature Extraction:

- Content features - punctuation and special characters are removed. Song lyrics are then vectorized using the TF-IDF technique. English stop words are also removed.

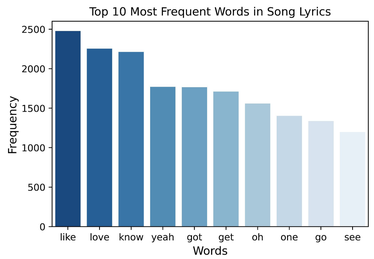


**Figure 2** – Top words found in the content feature set.

- Style features - uses the same protocol as content features. Only stop words are kept.

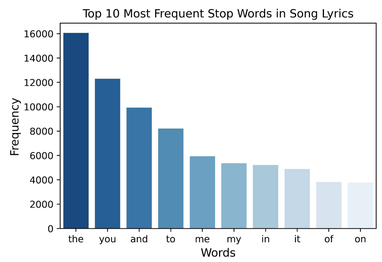


**Figure 3** – Top words found in the style feature set.

### Machine Learning Models:

2 Multi-label models are explored for this analysis. Similar to multi-class classification, the multi-label task has more than 2 classes. Unlike multi-class, one sample (or song) can be assigned more than one class (emotion)

- scikit-learn’s Logistic Regression and Naive Bayes classifiers are wrapped in a one-vs-rest classifier
- One classifier is fit per class. For each classifier, the class is fitted against all the other classes.

## Results

### Baseline:

ZeroR (or Zero Rate) Classifier is used as the baseline. ZeroR always classifies to the largest class - in other words, trivially predicting the most-frequent class.

For each emotion, we find the majority class (for example: ‘amazement’ (1) or ‘not amazement’ (0) and classify songs according to this majority. The majority class for each emotion was 0. In other words, not the emotion.

### Single-Emotion Classification:

| Emotion | Baseline | Content Features | | Style Features | |
|---|---|---|---|---|---|
| | | Logistic Regression | Naïve Bayes | Logistic Regression | Naïve Bayes |
| **amazement** | 79.40% | 78.28% | 78.28% | 78.28% | 78.28% |
| **calmness** | 78.45% | 75.52% | 75.52% | 75.52% | 75.52% |
| **joyful** | 69.91% | 71.38% | **72.41%** | 70.00% | 70.00% |
| **nostalgia** | 62.24% | 60.00% | 61.38% | 57.59% | 57.59% |
| **power** | 58.88% | 72.07% | **73.45%** | 70.34% | 66.90% |
| **sadness** | 50.52% | 71.72% | 67.59% | **72.07%** | 71.38% |
| **solemnity** | 67.50% | 67.24% | 67.93% | 67.24% | 67.24% |
| **tenderness** | 57.07% | 69.31% | **74.48%** | 62.41% | 57.93% |
| **tension** | 52.33% | **65.86%** | 63.45% | 61.72% | 60.34% |

**Figure 4** – Performance metrics per emotion. Bold cells indicate models with performance greater than Baseline

Models outperformed the baseline for the following emotions: **joyful, power, sadness, tenderness.**

- Models trained on content features tend to perform better than the baseline for these emotions - specifically Naive Bayes models

### Multi-Emotion Classification:

| MODEL | FEATURES | ACCURACY | F1 SCORE |
|---|---|---|---|
| Logistic Regression | Content | 6.89% | 0.4471 |
| Naive Bayes | Content | 5.52% | 0.3506 |
| Logistic Regression | Style | 3.10% | 0.4169 |
| Naive Bayes | Style | 0.34% | 0.2427 |

**Figure 5** – Model performance in multi-emotion classification.

All models yielded very low accuracies and performed worse than the baseline. However, the results from single-emotion classification resembled the results in multi-emotion classification, where accuracy was higher for models trained on content features as opposed to style features.

F1-score is used as a measure of a model's accuracy on a dataset. The F1 scores for content features are greater than those of style features, further confirming that models trained on content features perform better than style features.

## Discussion

Going into this project, we hypothesized that content-based feature sets would be more important in accurately classifying multiple emotions within a lyrical dataset. Our findings reveal that is not the case, and that style-based feature sets tend to more accurately classify melancholic emotions like 'sadness' or 'solemnity', which may indicate that the multi-emotion classification problems requires a more complex model.

While model performance in single-emotion classification (Table 2) is comparable to more sophisticated models, model performance is extremely weak in multi-emotion classification (Table 1). This may be representative of a dataset with lower inter-annotator agreement, where annotators may not label multiple emotions in a song in a consistent manner with other annotators, whereas the problem is easier in single-emotion labeling.

### Limitations

We also found that developing a style-based feature set from song lyric data is difficult to accomplish, given the arbitrary nature of transcribing punctuation from artists. Transcriptions can vary from artist to artist with little consistency, which underscores the potential importance of audio data[3] as opposed to text data in developing a style-based feature set for multi-emotion classification.

### Future Direction

Given the aforementioned limitations of extracting style-based features from lyric data, feature extraction from audio data in conjunction with lyric data is worth exploring for improved model performance. Additionally, given that the original dataset only contains song lyrics from four genres, a dataset with genres like a wider range of genres like pop or jazz, may allow for more information that allows for improved emotion classification.

## References

1. Edmonds, Darren, and João Sedoc. “Multi-Emotion Classification for Song Lyrics.” ACL Anthology, https://aclanthology.org/2021.wassa-1.24/.
2. Sari, Yunita, et al. “Topic or Style? Exploring the Most Useful Features for Authorship Attribution.” ACL Anthology, https://aclanthology.org/C18-1029/.
3. arXiv:1506.05012
4. Imdiptanu. “Imdiptanu/Lyrics-Emotion-Detection: Single-Label and Multi-Label Classifiers to Detect Emotions in Lyrics Achieved 0.65 and 0.82 F1 Scores Respectively.” GitHub, https://github.com/imdiptanu/lyrics-emotion-detection.