Comp 700 project proposal

Developing a machine learning model for automatic song lyric analysis and interpretation

Aryan Madhanjith (220005624)

Supervisor : Luke Vorster

# Abstract

The goal of this project would be to create a system that can analyse song lyrics and generate interpretations that are consistent with human understanding. This could involve training a machine learning model on a large dataset of annotated lyrics, where the annotations represent the interpretations of human experts.

The model could then be used to automatically analyse and interpret new song lyrics, and the results could be compared to the interpretations of human experts to evaluate the accuracy and effectiveness of the system. This type of research could have applications in various fields, such as music analysis, recommendation systems, and sentiment analysis.

# Introduction and Background

What do these lyrics mean?

*"Strumming my pain with his fingers*

*Singing my life with his words*

*Killing me softly with his song*

*Killing me softly with his song*

*Telling my whole life with his words*

*Killing me softly with his song"*

These lyrics are from the song "Killing Me Softly With His Song" originally performed by Roberta Flack and later covered by many other artists. The song is about a woman who is deeply moved by a singer's performance.

The lyrics suggest that the singer has a powerful ability to evoke emotions in the woman, as if he is strumming her pain with his fingers and singing her life with his words. The repetition of the line "Killing me softly with his song" emphasizes the intensity of the emotional impact. The singer's words are so powerful that they are able to tell the story of the woman's whole life.

Overall, the lyrics convey a sense of vulnerability and raw emotion in the woman, who is being deeply affected by the singer's performance. The song is a testament to the power of music to touch the soul and evoke powerful emotions in the listener.

Another interpretation of the song "Killing Me Softly" by Roberta Flack (or the Fugees) could be that it describes the emotional impact of music on the listener.

The singer describes being moved by a musician who is playing and singing with such passion and emotion that it feels like he is "strumming [her] pain" and "singing [her] life" with his words. The musician's performance is so powerful that it feels like he is "killing [her] softly" with his song.

In this interpretation, the song is not about a romantic relationship, but rather about the transcendent power of music to connect with people on a deep emotional level. The listener is so moved by the musician's performance that she feels like she is being "killed" by the intensity of her emotions, yet at the same time she is grateful for the experience of being so deeply affected by the music.

This interpretation is supported by the fact that the song has been covered by many different artists in various genres over the years, demonstrating its enduring appeal and ability to connect with people in different ways.

There are various websites and forums where people discuss and interpret song lyrics. Some of these include Genius, SongMeanings, and Reddit's r/LyricInterpretations. These platforms can offer different perspectives and interpretations of the same lyrics, which can be useful in gaining a broader understanding of a song's meaning.

It's important to note that the interpretation of song lyrics can be subjective and can vary from person to person, depending on their own experiences, emotions, and cultural background. Therefore, while these platforms can be helpful in exploring different interpretations of a song, it's ultimately up to each individual listener to form their own understanding of the lyrics.

# Literature review

## A) LYRICS-BASED MUSIC GENRE CLASSIFICATION USING A HIERARCHICAL ATTENTION NETWORK

Song lyrics can be thought of as having a hierarchical structure: Lines compose a verse/bridge/chorus, which in turn combine to make up the song. This paper suggests using hierarchical attention network (HAN) to learn the importance of each hierarchical element is to the genre type. It used various models (the “main” being the HAN ones) and compared accuracies over a 117- and 20-genre music dataset. The lyrics were not adapted to a Bag of Words model and were instead kept intact.

The HAN models have beaten all previously reported lyrical-only genre classification model accuracies, except for the classification among 5 genres. The accuracies obtained ranged from 45-49% [1].

However, there is intrinsically more to a song than just its lyrics, viz: artist background, chords used, etc. Mayer and Rauber used lyric and audio features to gain the highest accuracy of 74.08% [2].

The HAN can be used as one of the models to be tested.

## B) LYRICS-BASED ANALYSIS AND CLASSIFICATION OF MUSIC

The researchers have designed features that model semantic and stylistic properties of lyrics at a much deeper level and show that these features can indeed be beneficial [3]. This is opposed to earlier papers, which use a bag-of-words model – thus it differs to the previous paper in that it used the lyrics as is. The features are as follows:

* VOCABULARY:
  + These features estimate the vocabulary richness (type-token ratio) and the use of non-standard words.
  + These include “uncommon” and “slang” words (words not commonly found in dictionaries).
* STYLE:
  + POS tags are used to represent syntactic structure. To reduce data sparseness, all tags are mapped to more general ones (V, N, ADV).
  + Various length features are also implemented (lines per song, tokens per song, tokens per line).
  + Rhyme structure was modelled to detect perfect and imperfect in-line and line final rhymes.
  + Repetitions of letters (“riiiiise”) or words (“money, money”) are common in lyrics and often caused by a mismatch between number of syllables and line meter but they can also be employed as a means for emphasis and indicating emotion. We collectively dub such repetitions echoisms.
  + In-line (slant) rhymes (“burning turning”, “where were we”) are also grouped under ‘echoisms’. Echoisms are computed by looking for words with letter repetitions or word sequences with a relatively high similarity (according to an edit distance measure). Frequencies per type (letter reduplication, word repetition) and sequence length (less or more than 3 words) are encoded.
* SEMANTICS:
  + Lyrics can vary widely with respect to the topics they mention and the images they evoke.
  + Instead of using a linguistic model of semantic fields, we opted to build on work in psychology and use the Regressive Imagery Dictionary (RID) to identify predominant concepts (“imageries”) in a text.
  + RID classifies words as belonging to the separate fields “conceptual thought” (abstract, logical, reality-oriented), “primordial thought” (associative, concrete, fantasy), and “emotion”.
  + For example, the imagery ‘Moral’ (conceptual) contains words such as “should”, “right”, and “virtue”. Whereas the imagery ‘sensation’ (primordial) contains “delicious”, “perceive”, and “glamour”.
  + Intuitively, it is not only important what is said but also how it is said and the RID seemed to capture both aspects well. We computed the dominant imageries for each text and encoded this information in the feature vector.
* ORIENTATION:
  + This dimension models how the song narrative (entities, events) is oriented with respect to the world.
  + A temporal dimension is encoded, i.e., whether the song mainly recounts past experiences or present/future ones, by representing the fraction of past tense verb forms to all verb forms as a feature.
  + A model of how “egocentric” a song is, was also formed using the pronoun frequencies for 1st, 2nd, 3rd singular and plural person. As derived features, the proportion of self-referencing pronouns (first person singular/plural) to non-self-referencing ones and the ratio of first-person singular pronouns to second person was also encoded. The former feature measures the degree of talking about oneself as opposed to talking about other people, the latter measures whether the “I” or the “you” carries more weight in an interpersonal relationship.
* SONG STRUCTURE:
  + Structural repetitions are characteristic of song texts.
  + These include repetitive structures, i.e., identical, or similar multi-line blocks that re-occur, typically but not always representing the chorus.
  + Personally, I think this feature can be included in the dataset: A particular line is either part of a verse, chorus, bridge, etc.

2 models were created, one that just utilizes n-grams (baseline), and another that uses all of the features above (extended). The results (f1-scores) are interesting in that the baseline model outperformed the extended one by 3% on average. However, a combined model (uses the n-grams and the defined features) performed 3.5% better than the baseline model. This indicates that both models capture some, albeit different, aspects of each song.

We can include features defined here on our dataset of lyrics and meanings. An investigation can be made into whether these features will have any impact on the song interpretation.

# c) EMOTION ANALYSIS OF SONGS BASED ON LYRICAL AND AUDIO FEATURES

It is highlighted that people often most people can connect with the words of a song better than its musical features. In most cases, the lyrics are what truly express the emotions associated with the music, while the musical aspects are constructed around the lyrical theme. The paper method is proposed to detect the emotion of a song based on its lyrical and audio features.

The lyric-based features include the calculated metrics of “Arousal” and “Valence”, which refer to the intensity of emotion and the type of the emotion in a song respectively. Both are measured on a spectrum (lower values indicate negative emotions). This is intuitively sound, as emotions are rarely black and white, and a single song may carry a variety of feelings. All the audio features were gathered from “The Echo Nest – an Online Music Intelligence Platform.” These include Beats Per Minute, Mode, Loudness, Danceability and Energy.

For their experiment, the “emotions” used as labels for the songs included: Calm, Energetic, Dance, Happy, Sad Romantic, Seductive, Hopeful and Angry. Their dataset consisted of the most popular songs gathered from “Last.fm” related to each of the aforementioned tags (the ability of the site to incorporate user tags has made it popular in the construction of many other music datasets).

The classification is done by applying feature weighting and stepwise threshold reduction on the k-Nearest Neighbours algorithm where multiple classes are assigned to each song. Accuracy was measured by comparing the classes assigned by the algorithm to the classes derived from social tags obtained from “last.fm” .

Their classifications achieved an accuracy of around 82-84% [4] but it is noted that some emotions/classes are inherently subjective and there may be difficult to classify correctly. They have overcome this by labelling a classification as incorrect if it assigns a conflicting tag: If prediction is “Energetic/Dance” and the actual class is “Calm” (the opposite), the classification is incorrect.

The emotions attached to a song can influence the interpretations, hence our experiments can make use of the emotion/s as input along with the actual lyric interpretations.

## D) INTERPRETING SONG LYRICS WITH AN AUDIO-INFORMED PRE-TRAINED LANGUAGE MODEL

This paper [5] proposes a “BART-fusion”, a novel model for generating lyric interpretations from lyrics and music audio that combines a large-scale pre-trained language model with an audio encoder. It is shown that including the audio and the lyrics representation to help the pre-trained language model understand the song from an audio perspective, while preserving the language model’s original generative performance based on the lyric interpretations.

Sequence-to-sequence tasks generally have one encoder and one decoder. In order to combine the audio and lyric interpretations, two encoders are used. The lyric encoder is built following the traditional architecture (lyrics are tokenized and positionally encoded). The audio encoder receives a part of a song and various features are extracted (yet also follows the traditional architecture):

Consider a diagram to illustrate its complexity:

A computer screen shot of a computer screen

Description automatically generated with low confidence

The other important take-away from this team of researchers would be their “Song Interpretation Dataset” [8]. It contains the audio, song lyrics and various human/user interpretations. It also includes the number of votes for each interpretation (indicating how many people agree with said interpretation).

The paper uses various versions of the dataset based on voting (one keeps interpretations having only positive votes, and another uses all interpretations). These datasets were used on their BERT-fusion model and a “regular” BERT model. ROUGE, METEOR and BERT-Score to evaluate the generated interpretations. The results indicate that their combined audio and lyric model generally performs better across all versions of the dataset.

It is shown that audio features can be used in conjunction with the lyric interpretations. Our models can also follow this structure of combination to achieve better results.

## E) ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS

Making a language model larger to facilitate its improvement is often not feasible, due to long training times, limited memory and in some cases, the degradation in performance when the number of parameters is increased.

ALBERT [6] seeks to improve on the traditional BERT model using the following techniques:

* “Factorized embedding parameterization”, where the size of the hidden layers is separated from the size of vocabulary embeddings by transforming the large vocabulary-embedding matrix into two small matrices.
* “Cross-layer parameter sharing”, which prevents the number of parameters from growing with the depth of the network.
* Self-supervised loss for sentence-order prediction, which specifically addresses BERT’s limitations with regard to inter-sentence coherence.

This improved model has yielded x1.7 faster training times using x18 fewer parameters. Although it does achieve slightly lower performance-wise compared to the original BERT model.

For our purposes, it will be worthwhile to consider using ALBERT based on the current limited resources and restraints placed by Google Colab.

## F) Genre of music and lyrical content: expectation effects

This is not a computer science research paper, but we want to know, for purposes related to our dataset if the genre of a song has any influence on the interpretation of its lyrics.

The study [7] was performed on 160 people. They were asked to give an interpretation of lyrics passages (that were selected by researchers as prosocial and antisocial) presented as four different music genres including heavy metal rap pop and country (which have different dispositions toward the inclusion of prosocial and antisocial lyrics).

It was found via a questionnaire that the labelled genre indeed influenced the impact of the lyrics. Specifically, lyrics labelled as heavy metal or rap would be perceived as more antisocial than the same lyrics labelled as country or pop.

Thus, in our dataset, we must also include the genre of a song alongside the interpretation.

# Proposed methods and techniques

The research methodology will be largely empirical, as it involves developing and testing language models, as well as evaluating them based on their performance metrics. They will also be subjected to training on different versions of the dataset to observe how they perform with and without other elements/features of the data.

To a small degree, the research is also theoretical in nature, as we are further developing the language models to include the audio features in the encoding stage. This is a relatively unexplored avenue as I have only found one research paper/team that has developed this hybrid encoder.

The Song Interpretation Dataset [8] will be used, which contains audio excerpts from 27,834 songs (30 seconds long), the corresponding music metadata, about 490,000 human interpretations of the given lyric text for that audio clip, and the number of votes given for each of these user interpretations.

The average length of the interpretations is 97 words. Compared to another popular music site, Genius Lyrics, whose interpretations are often one or two sentences (most of the time, referencing contemporary events based on the artist/s), the Song Interpretation Dataset is a better choice as it provides more insight.

Music in the dataset covers various genres, of which the top 5 are: Rock (11,626), Pop (6,071), Metal (2,516), Electronic (2,213) and Folk (1,760). It is important to distinguish these classes, as interpretations have proven to be influenced by them.

The audio features can be gathered from a music API known as “The Echo Nest”.

### Evaluations

As the nature of interpretations is subjective, we cannot use the typical metrics (namely accuracy). Although, several metrics have been developed to measure the performance of text summarizers, which can also be used for our tests:

* Recall-Oriented Understudy for Gisting Evaluation/ROUGE [9]:
  + It works by comparing an automatically produced summary or translation (in our case, the model interpretation) against a set of reference summaries which are usually human made (human interpretations).
  + It simply involves calculating the precision and recall:
    - Recall =
    - Precision =
    - where the overlapping words refers to the words appearing in both the model output and the human interpretation
  + We can clearly see why this is more suited to text summarization, but it can still be used to show that the model interpretations carry the same information as the human ones.
* Metric for Evaluation of Translation with Explicit ORdering [10]:
  + A metric used in machine translation, which also works based on precision and recall.
  + Based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations.
  + It can work with sentences in the same language, but in this case, the score is a representation of the similarities between the sentences.
  + It is also capable of considering multiple references for one prediction (which will be useful as our dataset contains many user interpretations of lyrics).
* BERT-score [11]:
  + It uses the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
  + It is best used when measuring the correlation between “human-judgement” and machine level evaluation.

# Time frame

|  |  |
| --- | --- |
| Date | Event |
| 24th March | Register project with module co-ordinator |
| 5th May | Present initial proposal to supervisor |
| 30th June | Submit proposal |
| 1st July – 30th July | Begin developing models for testing |
| 1st July – 30th July | Start drafting research paper\* |
| 31st July | Have at least one model working |
| 1st August – 28th September | Finalize project and research paper |
| 29th September | Present research paper draft to supervisor |
| 27th October | Submit project (software and research paper) |
| 9th – 10th November | Oral presentations of project |

\* It may be beneficial to do so during the development phase rather than after it.

# References

[1] A . Tsaptsinos. “Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network”, *18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017

[2] Mayer and A. Rauber. Musical genre classification by ensembles of audio and lyrics features. In *ISMIR*, pages 675–680, 2011

[3] M. Fell, C. Sporleder, “Lyrics-based analysis and classification of music”, *the 25th International Conference on Computational Linguistics: Technical Papers*, pages 620–631, Dublin, Ireland, August 23-29, 2014

[4] A. Jamdar, J. Abraham , K. Khanna, R. Dubey, “Emotion analysis of songs based on lyrical and audio features”, *International Journal of Artificial Intelligence & Applications (IJAIA) Vol. 6*, No. 3, May 2015

[5] Y. Zhang, J. Jiang, G. Xia, S. Dixon, “Interpreting song lyrics with an audio-informed pre-trained language model”, in Proc. of the 23rd Int. Society for Music Information Retrieval Conf., Bengaluru, India, 2022

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, “ALBERT: a lite BERT for self-supervised learning of language representations”, *ICLR*, 2020

[7] M. E. Ballard, D. G. Bazzini, A. R. Dodson, “Genre of music and lyrical content: expectation effects”, *The Journal of Genetic Psychology*, 160(4): 476-487. (Dec 1999) Published by Taylor & Francis (ISSN: 1940-0896)

[8] Y. Zhang, J. Jiang, G. Xia, S. Dixon. (2022). Song Interpretation Dataset (1.2) [Data set]. Zenodo.

[9] C. Y. Lin, “ROUGE: A package for automatic evaluation of summaries”. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics, 2004.

[10] S. Banerjee A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”, In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65-72. 2005.

[11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, “BERTSCORE: evaluating text generation with BERT”, *In ICLR,* 2020