**ANL488 PROJECT PROPOSAL**

**The Evolution of Popular Tunes**

A picture containing icon

Description automatically generated

**Submitted by**

**Name: Tan Li Lin**

**PI Number: W1882296**

**SCHOOL OF BUSINESS**

**Singapore University of Social Sciences**

**Presented to Singapore University of Social**

**Sciences in partial fulfilment of the**

**requirements for the Degree of Bachelor of Science**

**in Business Analytics**

**2021**

**Table of Contents**

[Chapter 1 Introduction 3](#_Toc79995561)

[Chapter 2 Literature Review 4](#_Toc79995562)

[Chapter 3 Data Understanding and Preparation 7](#_Toc79995563)

[Chapter 4 Proposed Modeling and Evaluation 21](#_Toc79995564)

[Chapter 5 Proposed Schedule 23](#_Toc79995565)

[References 27](#_Toc79995566)

# **Chapter 1 Introduction**

Music is seen as an effective medium that fosters nonverbal communication, allows meaning to be conveyed, and creates national identities. In addition, listening to music has significant therapeutic results which can reduce anxiety, promote relaxation as well as improve an individual's quality of life (Music Magic, 2008). Therefore, music is perceived as an important constituent in our everyday life, be it for music creation, performance, pleasure, or emotional response (Galindo, 2003).

Over the centuries, it is apparent that music has changed along with society regardless of the tunes or lyrics used in each piece of music. In the aspect of tunes, in the earlier, it resembled closely to the nature of ambience, whereas in the latter, more musical instruments are introduced which produced relatively sophisticated tunes (Henry, 2018).

On the other hand, songwriters convey their thoughts through lyrics to enable listeners to view and relate things from their perspective (Winston, 2017). As different songwriters express themselves differently, their state of mind differs when lyrics are composed. Thus, these lyrics can be used to provide sentiment insights.

Lyrics are in a form of textual data which require text mining techniques to clean and process to extract relevant and useful information. In particular, it seeks to uncover the sentiments of the text to determine the state of mind of songwriters when they composed music.

This study goes about reviewing the various approaches to analyze textual data. It then aims to extend these techniques to analyze the sentiments of bilingual popular music through lyrics that was composed in the year between 1970 to 2020. In addition, it then identifies and classifies the patterns of musical sentiments over the years.

# **Chapter 2 Literature Review**

As lyrics are penned from the songwriter's thoughts, therefore, they are considered as user-generated content (UGC) (Barman, Dahekar, Anshuman, & Awekar, 2019). With the increasing interest in studying the thoughts behind UGC, sentiment analysis is a technique to classify the polarity by employing machine learning techniques, like Naïve Bayes (NB) Classifier and Support Vector Machine (SVM) (MonkeyLearn1, n.d.).

In the study by Hu, Downie, and Ehmann (2009), they explored how lyrics can guide in classifying the mood of the music. With that, they gathered approximately 21,000 music from online lyrics databases and social tags from last.fm. Nonetheless, following data exploration and preparation of eliminating insignificant tags that are non-affective, judgmental, and have ambiguous meanings, as well as integrating synonym tags, the finalized dataset comprised of 5,585 pieces of music and 18 mood categories. As the accuracy data are rarely normally distributed, they adopted non-parametric Friedman's ANOVA test to determine if there was a significant difference in the performance. Furthermore, they adopted SVM as the classifier model for its superior performance in text categorization and Music Information Retrieval (MIR) tasks. Thus, they built models to test the accuracy of the categories as well as the performance of combined features of both audio and lyrics. With multiple models built to test the accuracy of categories, Bag-of-Words (BOW) with stemming and tf-idf weighting achieved a higher average accuracy of 0.6043. As a result, this model was used to examine the following model of analyzing the performance of combined features, which concluded that combined features did enhance the performance for the majority of the categories, but lyrics-only can outperform audio-only if it was classified under the relevant mood category.

However, due to the lack of appropriate techniques for analyzing multilingual data, most research studies focused mainly on the common language, English. Yan, He, Shen, and Tang (2014) did the exceptions by gathering a total of 4,000 bilingual review comments from Facebook, Twitter, Tianya forum, and Weixin, on a popular movie to assess the suitability of proposed models, SVM and N-gram, for sentiment analysis. The data comprised an equal proportion of positive and negative comments for both English and Chinese respectively. Of all comments, 80% of it for each respective language are set to train the models and the remaining to test the trained models. Before training the models, Yan et al. (2014) made a few significant pointers that there are various approaches to segment Chinese sentences, and in the language of Chinese, it has a distinct way of expressing emotions. Hence, the sentiment analysis technique that was developed for English might not be suitable to deal with Chinese directly. Therefore, they adopted a widely used open-source application, IKAnalyzer, to perform segmentation for these comments. The trained models suggested that SVM performed better as compared to N-gram with higher accuracy of 98.90% and 82.42% respectively. Hence, the study concluded that SVM was a more appropriate model to analyze bilingual textual data although it highlights that Chinese achieved a slightly lower accuracy of 85% which could likely be because Chinese segmentation is not entirely accurate.

Relatively closer to the following study, Napier, and Shamir (2018) studied how lyrics changed from the 1950s to the present by employing digital humanities and data science techniques, and then perform quantitative analysis to quantify these changes. They gathered a total of 6,085 pieces of pop music containing lyrics from Billboard Hot 100 songs, from 1951 to 2016. They adopted IBM Watson Tone Analyzer to evaluate lyrics for the tone to determine the musical sentiments conveyed. Extensively, Tone Analyzer examines the combination of distinct words and tones using SVM, with a one-vs-rest approach to extends SVM to more than two classes. Furthermore, the choice of words used in lyrics provides significant information about the tone and songwriter's personality for the computer to evaluate. In addition, two tests of Pearson correlation and linear regression were performed with the use of averaged tone scores, to determine the correlation between the tone in lyrics and the year composed. This study concluded that the tone of music lyrics has evolved significantly over time, along with social values as conveyed in the shifts in mainstream popular music.

# **Chapter 3 Data Understanding and Preparation**

Assumptions were established during the data collection process. A list of bilingual popular music from 1970 to 2020 is being consolidated under the assumption that music becomes popular in the year it is released. As a result, in this case, both English and Chinese popular music will be used to study the musical sentiments over time through lyrics. The consolidated list of music lyrics is manually retrieved from the internet and entered into excel. Specifically, depending on the results of Google search, English lyrics are retrieved from either https://www.lyricfind.com/, https://www.musixmatch.com/, or https://www.lyrics.com/, whereas Chinese lyrics are retrieved from a Chinese portal, namely https://baike.baidu.com/, which functions as a search engine in Mainland China.

Table 1 – Variable Description

|  |  |  |  |
| --- | --- | --- | --- |
| S/N | Variable | \* Type | Description |
| 1 | Year | Number | The year when music is released |
| 2 | Music Title | String | The title of the music |
| 3 | Artist | String | Someone who composes, performs, and releases music |
| 4 | Genre | String | The genre of the music |
| 5 | Duration | String | The length of the music |
| 6 | # of views | Number | \*\* The number of views in YouTube for each music |
| 7 | Lyrics | String | The lyrics of the music |
| 8 | Mood | String | The mood that the music creates |
| 9 | Emotions Profile | String | The overview of musical sentiment |

\* Type column is retrieved from Tableau for data exploration

\*\* The maximum number of watched views retrieved in YouTube for each specific music.

Text

Description automatically generated

Figure 1 – Information on English lyrics dataset

Text

Description automatically generated

Figure 2 – Information on Chinese lyrics dataset

As illustrated in Figures 1 and 2, a total of 100 lyrics are obtained with an equal proportion of English and Chinese music according to the nine variables of 'Year', 'Music Title', 'Artist', 'Genre', 'Duration', '# of views', 'Lyrics', 'Mood' and 'Emotions Profile'.

Graphical user interface, text, application

Description automatically generated

Figure 3 – Sample data for English lyrics

A picture containing application

Description automatically generated

Figure 4 – Sample data for Chinese lyrics

Chart, bubble chart

Description automatically generated

Figure 5 – No. of English popular music by Year

Chart, bubble chart

Description automatically generated

Figure 6 – No. of Chinese popular music by Year

In Figure 5, it depicts that the most popular English music was composed in the following years: 1970, 1971, 1981, 1991, and 2011 with a maximum of three pieces of music, whereas in Figure 6, it depicts that most popular Chinese music was composed in the year 1979 with a maximum of five pieces of music, followed by 2017 with four pieces of music.

A picture containing text

Description automatically generated

Figure 7 – Word cloud of all English lyrics

A picture containing text

Description automatically generated

Figure 8 - – Word cloud of the genre of English music

A picture containing diagram

Description automatically generated

Figure 9 – Word cloud of all Chinese lyrics

A picture containing logo

Description automatically generated

Figure 10 – Word cloud of the genre of Chinese music

Word cloud is an indication of word frequency, with larger font sizes denoting higher frequency. To construct Chinese characters word cloud, 'Jieba' package is used as it is known to be the best Python module for Chinese word segmentation (Develop Paper, 2021). Figures 7 and 9 illustrate the frequency of words used in lyrics for both English and Chinese music respectively. Whereas Figures 8 and 10 illustrate the genre of popular music, with 'Pop' being the most popular followed by 'Rock'.

Chart

Description automatically generated with low confidence

Figure 11 –No. of Views by Year Range (English Music)

Text, application

Description automatically generated

Figure 12 – No. of Views by Year Range (Chinese Music)

Figures 11 and 12 illustrate the total number of views by the year range of '1970 - 1980', '1980 - 1990', '1990 - 2000', '2000 - 2010' and '2010 - 2020'. As illustrated, there are more watched views between the years 2010 - 2020. However, watched views is an inaccurate measure as the advancement of recording technology occurred in the 20th century where listeners have access to a vast variety of music 24/7, at the flick of a switch (Music Magic, 2008).

Chart, bar chart

Description automatically generated

Figure 13 – Count of Emotions Profile (English Music)

Chart, bar chart

Description automatically generated

Figure 14 – Count of Emotions Profile (Chinese Music)

Figures 13 and 14 illustrate the overall count of emotions profile for each respective language, which include balanced, negative, positive, and variable. Balanced signifies that music conveys a neutral sentiment whereas variable signifies that music conveys a positive or negative sentiment which could determine by individuals. The emotions profile represents the sentiment of the music, as determined by CYANITE, an online platform that visualizes music metadata (CYANITE, n.d.). It appears that English and Chinese music convey opposing sentiments, with English popular music conveying a more positive vibe and Chinese popular music conveying a more negative vibe. It is important to note that denoting a negative emotion profile does not imply that it delivers negativity, but rather a sorrowful feeling.

Moving on to data quality, there was no concerns as it is manually retrieved from the internet with the necessary variables. However, data preparation is required to better grasp the musical sentiment through lyrics.

Text pre-processing will be performed at the data preparation stage. However, the variables 'Duration' and '# of views' do not add significant value to the upcoming models for sentiment analysis, thus, they will be eliminated.

A picture containing text

Description automatically generated

Figure 15 – After elimination of insignificant variables (English music dataset)

Table

Description automatically generated with low confidence

Figure 16 – After elimination of insignificant variables (Chinese music dataset)

In this study, text analysis will perform on a document level where one music lyric represents a document. Steps taken in each text pre-processing stages will be accounted for in Table 2 and Table 3.

Table 2 – Text pre-processing for English Lyrics

|  |
| --- |
| **Text Pre-Processing for English Language** |
| **Data Cleaning:** |
| 1. Tranformed text to lowercase and duplicate it to a new column, cleaned\_text.    Figure 17 – Example after case normalization  2. Imported contradiction dictionary. For example, 'what'll' will be transformed into 'what will'. This transformation replaced values in the column, cleaned\_text.    Figure 18 – Sample after contractions  3. Replaced column, cleaned\_text, values after removing punctuations and replacing new lines to space.    Figure 19 – Sample after removing of punctuations, new line |
| **Tokenization + Part of Speech (POS) tagging + Stopwords:** |
| Tokenization is the split of sentences into individual words, while POS identifies the relevant word class, such as a noun or a verb. Common stopwords are also eliminated from the results after tokenization and POS tagging. In addition, a list of words with insignificant values is added to the list of stopwords as well.    Figure 20 – Tokenization and POS tagging    Figure 21 – After stop words on the results of tokenization and POS tagging  As seen from the above figures, stop words eliminate common words such as 'I', "You", 'Him', etc. |
| **Lemmatization:** |
| Lemmatization is performed to map words to their root term. Stopwords are deployed under the lemmatization process as lemmatization is done based on the tokenized lyrics that were yet to process on POS tagging and stopwords which was mentioned above.    Figure 22 – Lemmatization of words to its root term  For example, as shown in Figure 22, the word 'womans' is mapped to 'woman', and the word 'times' in the original text is mapped to the word 'time'. |

Table 3 – Text pre-processing for Chinese Lyrics

|  |
| --- |
| **Text Pre-Processing for Chinese Language** |
| **Data Cleaning:** |
| 1. Case normalization is performed as certain lyrics contain English lyrics. Tranformed text to lowercase and duplicate it to a new column, cleaned\_text.    Figure 23 – Example after case normalization  2. Replaced column, cleaned\_text, values after removing punctuations and replacing new lines to space.    Figure 24 – Sample after removing of punctuations, new line |

|  |
| --- |
| **Tokenization & Part of Speech (POS) tagging** |
| As mentioned above, tokenization is to split sentences into individual words, and POS associates the relevant word class. Therefore, 'Jieba' module is used for both tokenization and POS tagging of the word.    Figure 25 – Tokenization and POS tagging |
| Stop words and the lemmatization process are not used in Chinese text pre-processing. The reason being is when stop words are applied, the meaning of Chinese sentences changes. In addition, Chinese words do not contain any tenses, hence lemmatization is not applied. |

Table

Description automatically generated

Figure 26 – TF-IDF of English music lyrics

Table

Description automatically generated with medium confidence

Figure 27 – TF-IDF of English music lyrics

Table 4 – Number of rows and Columns generated for TF-IDF

|  |  |  |
| --- | --- | --- |
|  | **Rows** | **Columns** |
| **TF-IDF (English)** | 50 | 65 |
| **TF-IDF (Chinese)** | 50 | 2266 |

Term frequency-inverse document frequency (TD-IDF) is a statistical measure that determines how relevant a word is to a document. This is achieved by multiplying two metrics: the number of times a word appears in a document, and the word's inverse document frequency over a set of documents. Both Figures 26 and 27 illustrate the TF-IDF of all words for both English and Chinese lyrics, while Table 4 illustrates the number of words after the pre-processing of tokenization.

# **Chapter 4 Proposed Modeling and Evaluation**

Positive, negative, and neutral sentiment are the three types of sentiment. As a result, the techniques mentioned below will be applied to determine the sentiment of the lyrics.

The sentiment score and tone of bilingual popular music will be analyzed using IBM WTA. It is a tool that uses linguistic analysis to detect emotions and linguistic tones in written text (IBM1, 2020). However, this program is only available to analyze texts that are either English or French (IBM2, 2020). Therefore, it will only be used to analyze the sentiment score and tone for English lyrics.

Due to IBM WTA's inability to analyze through Chinese language and the scarcity of online resources, an online AI Builder, sentiment analysis prebuilt model, from Microsoft PowerApps will be used. It is a tool that detects sentiment in text data and returns emotions and probability scores (Microsoft, 2019). However, a manual extraction of the scores into excel is required.

As the report is about classification, the following models will be the proposed: SVM and NB, to be implemented after the computation of sentiments. The data will be partitioned into a 70/30 rule with 70% of the data to train to build the model and the remaining 30% to test the model, for each language.

Specifically, SVM and NB classifiers do text classification which is aligned to the result of the report. SVM is the coordinates of individual observation that separates classes for easy identification of SVM (Ray, 2017) whereas the NB model is built based on Bayes’ Theorem to compute the conditional probability of occurrence of two events depending on the probabilities of occurrence of each event. Thus, assisting in the classification of text by predicting the likelihood of text being placed in the respective categories (MonkeyLearn2, n.d.).

The models will then be evaluated based on the accuracy, precision, recall, and F1 score. Accuracy is a widely used measure to evaluate models, but it might not be a reliable indicator when classes are unbalanced. Precision determines the hit rate that is classified correctly while recall determines the wrongly classified ones. Lastly, the F1-score is used as a measure to determine the balance between precision and recall (Singh, 2019).

Additionally, visualization will be built to illustrate the musical sentiment over the years.

# **Chapter 5 Proposed Schedule**

|  |  |  |
| --- | --- | --- |
| Project Milestone | | |
| Start Date - End Date | **Milestone** | **Duration** |
| 18-May-21 to 27-May-21 | Submit intention survey | 10 days |
| 1-Jun-21 | ANL488 Pre-briefing | 1 day |
| 01-Jun-21 to 10-Jun-21 | Topic selection | 10 days |
| 18-Jun-21 | Topic + Supervisor allocation | 1 day |
| 19-Jun-21 to 02-Jul-21 | **Work on Project Proposal (draft)**   * Understand the project description * Explore possible dataset | 14 days |
| 06-Jul-21 | **Pre-course meeting with supervisor**   * The expectation of the project | 1 day |
| 07-Jul-21 to 22-Jul-21 | **Work on Project Proposal (draft)**   * Consolidated a list of data to be used * Craft a problem statement for the project | 16 days |

|  |  |  |
| --- | --- | --- |
| 23-Jul-21 | 1st meeting with supervisor   * Discussion on work progress | 1 day |
| 26-Jul-21 | First seminar | 1 day |
| 27-Jul-21 to 05-Aug-21 | **Work on Project Proposal**   * Extract necessary data (of various variables) from open source (online) * Research on relevant articles with relevant techniques * Introduction | 10 days |
| 6-Aug-21 | **2nd meeting with supervisor**   * Discussion on work progress | 1 day |
| 07-Aug-21 to 15-Aug-21 | **Work on Project Proposal**   * Literature Review * Data Understanding and Preparation * Proposed Modeling and Evaluation * Proposed Schedule | 9 days |
| 16-Aug-21 | Project Proposal submission | 1 day |
| 17-Aug-21 to 09-Sep-21 | **Work on Final Report**   * Revise Project Proposal according to feedback received from supervisor * Modeling * Evaluation * Recommendations / Conclusion | 24 days |
| 10-Sep-21 | **3rd meeting with supervisor**   * Discussion on work progress | 1 day |
| 11-Sep-21 to 19-Sep-21 | **Work on Final Report**   * Fine-tune modeling and evaluation based on feedback given by supervisor from the previous meeting | 9 days |
| 20-Sep-21 to 25-Sep-21 | Oral Presentation | 6 days |
| 29-Sept-21 | **5th meeting with supervisor**   * Feedbacks from supervisor * Discussion on work progress | 1 day |

|  |  |  |
| --- | --- | --- |
| 30-Sep-21 to 14-Oct-21 | Work on Final Report   * Fine-tune report based on feedback received from oral presentation * Prepare Final Report | 15 days |
| 15-Oct-21 | **5th meeting with supervisor**   * Discussion on work progress | 1 day |
| 16-Oct-21 to 07-Nov-21 | **Work on Final Report**   * Prepare and finalize Final Report | 23 days |
| 08-Nov-21 | Final Report Submission | 1 day |

# **References**

Barman, M. P., Dahekar, K., Anshuman, A., &Awekar, A. (2019). It’s only Words and Words Are All I Have. *Lecture Notes in Computer Science*, 30-36. doi: 10.1007/978-3-030-15719-7\_4.

CYANITE (n.d.). *About*. Retrieved August 12, 2021, from <https://cyanite.ai/about/>.

Develop Paper (2021). *Detailed use in Chinese word segmentation based on Jieba package in Python.* Retrieved August 11, 2021, from <https://developpaper.com/detailed-use-in-chinese-word-segmentation-based-on-jieba-package-in-python/>.

Galindo, G. (2003). *The Importance of Music in Our Society.* Retrieved August 4, 2021, from <https://www.gilbertgalindo.com/importanceofmusic>.

Henry (2018). *How the Sound of Music Has Changed Over the Years.* Retrieved August 4, 2021, from <https://sonicspace.org/how-the-sound-of-music-has-changed-over-the-years/>.

Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric Text Mining in Music Mood Classification. *Proceedings of the International Society for Music Information Retrieval Conference*, 411 – 416.

IBM1 (2020). *About.* Retrieved August 14, 2021, from <https://cloud.ibm.com/docs/tone-analyzer?topic=tone-analyzer-about>.

IBM2 (2021). *Using the general-purpose endpoint*. Retrieved August 14, 2021, from <https://cloud.ibm.com/docs/tone-analyzer?topic=tone-analyzer-utgpe>.

Microsoft (2019). *Sentiment analysis prebuilt model.* Retrieved August 16, 2021, from <https://docs.microsoft.com/en-us/ai-builder/prebuilt-sentiment-analysis>.

MonkeyLearn1 (n.d.). *Sentiment Analysis: A Definitive Guide.* Retrieved August 4, 2021, from <https://monkeylearn.com/sentiment-analysis/>.

MonkeyLean2 (n.d.). *Text Classification Using Naïve Bayes*. Retrieved August 15, 2021, from <https://monkeylearn.com/text-classification-naive-bayes/>.

Music Magic (2008). *The Powerful Role of Music in society*. Retrieved August 4, 2021, from <https://musicmagic.wordpress.com/2008/07/10/music-in-society/>.

Napier, K. & Shamir, L. (2018). Quantitative Sentiment Analysis of Lyrics in Popular Music. *Journal of Popular in Music Studies*, 30(4), 161-176. doi: 10.1525/jpms.2018.300411.

Ray, S. (2017). *Understanding Support Vector Machine(SVM) algorithm from examples (along with code).* Retrieved August 15, 2021, from <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>.

Singh, B. (2019). *Evaluation Metrics for Machine Learning Models.* Retrieved August 15, 2021, from <https://heartbeat.fritz.ai/evaluation-metrics-for-machine-learning-models-d42138496366>.

Winston, C. (2017). *Why Do Lyrics Matter?* Retrieved August 4, 2021, from <https://www.nrgrecording.com/post/why-do-lyrics-matter>.

Yan, G., He, W., Shen, J., & Tang, C. (2014). A bilingual approach for conducting Chinese and English social media sentiment analysis. *Computer Networks*, 75, 491-503. doi: 10.1016/j.comnet.2014.08.021.