ANL488 PROJECT PROPOSAL

The Evolution of Popular Tunes

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 4

Chapter 2 Literature Review 6

Chapter 3 Data Understanding and Preparation 9

Chapter 4 Proposed Modeling and Evaluation 23

Chapter 5 Proposed Schedule 25

References 29

Here are my main comments

1) ***Topic Formulation***: A good topic choice; it is correct that often sentiment analysis is focused on English so looking at how music evolves in another language is very interesting for sure
2) ***Literature Review***: Your literature was highly technical and was the right length. You "hit hard" with the right papers, facts etc. My one comment would be that it stood like 3 topics (mood, bilingual, audiation/tone) instead of 1 singular topic. You could have easily addressed this with a "harmonizing" paragraph.
3) ***Data Understanding***: I like that you state the assumptions upfront. That's important in any analytics topic because as we know data is imperfect. I will also be honest and tell you now that I did not like how you wrote this section because you didn't think of the reader; you just put pictures in places where text would have been better to EXPLAIN your ideas. Another thing that is lacking is how the languages differ – you are assuming that the reader is already familiar with the Chinese language sentence structure, for example
4) ***Proposed Modeling***: No major issues with the proposed method; I would encourage you to develop some baseline comparators since you are using 2 different software/methods?
5) ***Overall Presentation***: It was a interesting topic, and I can tell that you have done quite a bit of research. So I really am impressed by your effort.

My main feedback is that it didn't read smoothly; you have a really cool idea and I am super excited to see the results. However, it sounds mechanical when you write it; like Step A – do this, Step B – do that….this statement really says what you are trying to so "This study concluded that the XXXXX of music lyrics has evolved significantly over time, along with social values as conveyed in the shifts in mainstream popular music." But now, you need to enhance this statement to say that you will be studying Chinese and English music to confirm this.

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.

Commented [MK1]: ok

Commented [MK2]: OK, I can see where you are going with this

Commented [MK3]: I think what you mean here is that lyrics are suitable for.....

Commented [MK4]: Ok this is your topic; you will first study text analysis techniques, then apply it to bilingual music. Its not clear here what is the questions you are addressing. You are just stating what you are going to do.....

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) (MonkeyLearn[1], 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.

Commented [MK5]: What mood categories? Is this something you will also study?

Commented [MK6]: What is this? Do you need to explain this?

Commented [MK7]: While this is an interesting result, it is discussing mood? I assume that this is something you want to apply to your bilingual topic?

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.

Commented [MK8]: This is interesting; you need to explain what IKAnalyzer is about a bit more. I guess this will be very relevant to your work.

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.

Commented [MK9]: So wahts the linkage from all your literature review with what you want to do? What are you challenging. Fixing addressing

Commented [MK10]: Ok good

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.

Commented [MK11]: good

Commented [MK12]: ok

Table 1 – Variable Description

Commented [MK13]: Never do this again, Li Lin

When you have table 1, figure 1, table 2, figure 2 etc, you need to reference it in your text and EXPLAIN whar each table means. It cannot be expected that I scroll through your list of tables/ figures and know what you are showing me!

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.

English Lyrics Dataset Information

```
eng.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 9 columns):
Year                50 non-null int64
Music Title         50 non-null object
Artist              50 non-null object
Genre               50 non-null object
Duration            50 non-null float64
# of views          50 non-null object
Lyrics              50 non-null object
Mood                50 non-null object
Emotions Profile    50 non-null object
dtypes: float64(1), int64(1), object(7)
memory usage: 3.6+ KB
```

Figure 1 – Information on English lyrics dataset

Commented [MK14]: How can your figure come before your text?

Chinese Lyrics Dataset Information

```
chi_lyrics.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 9 columns):
Year                50 non-null int64
Music Title         50 non-null object
Artist              50 non-null object
Genre               50 non-null object
Duration            50 non-null float64
# of views          50 non-null object
Lyrics              50 non-null object
Mood                50 non-null object
Emotions Profile    50 non-null object
dtypes: float64(1), int64(1), object(7)
memory usage: 3.6+ KB
```

Figure 2 – Information on Chinese lyrics dataset

Commented [MK15]: See above comment

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'.

	Year	Music Title	Artist	Genre	Duration	# of views	Lyrics	Mood	Emotions Profile
0	1977	Stayin' Alive	Bee Gees	Pop	4.09	594,869,283	Well, you can tell by the way I use my walk\nI...	Energetic, Happy, Uplifting	Positive
1	1970	Layla	Derek and the Dominos	Rock	8.01	140,997,541	What'll you do when you get lonely\nAnd nobody...	Energic, Epic	Balanced
2	1978	Y.M.C.A	Village People	Rock	4.01	245,509,328	Young man, there's no need to feel down\nI sai...	Energetic, Happy, Sexy, Uplifting	Positive

Figure 3 – Sample data for English lyrics

	Year	Music Title	Artist	Genre	Duration	# of views	Lyrics	Mood	Emotions Profile
0	1978	夜來香	邓丽君	Rock	3.20	604,434	那南风吹来清凉\n那夜莺啼声细唱\n月下的花儿都入梦\n只有那夜来香\n吐露着芬芳\n我爱这...	Chill, Happy	Positive
1	1977	月亮代表我的心	邓丽君	Latin	3.24	11,819,184	你问我爱你有多深 我爱你有几分\n我的情也真 我的爱也真\n月亮代表我的心\n你问我爱你有多...	Calm, Chill, Romantic, Sad, Ethereal	Negative
2	1979	甜蜜蜜	邓丽君	Pop	3.30	2,045,273	甜蜜蜜 你笑得甜蜜蜜\n好象花儿开在春风里\n开在春风里\n在哪里 在哪里见过你\n你的笑容...	Calm, Chill, Romantic, Sad, Ethereal	Balanced

Figure 4 – Sample data for Chinese lyrics

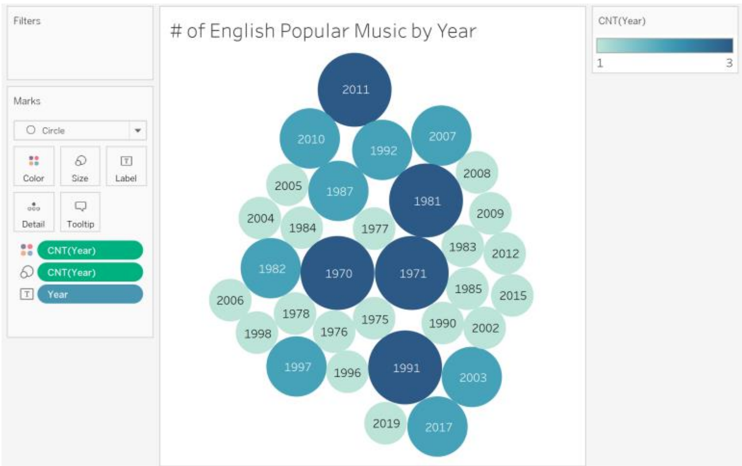

Figure 5 – No. of English popular music by Year

Commented [MK16]: Where is this discussed in the text? This is not art class. 😊

Commented [MK17]: See above

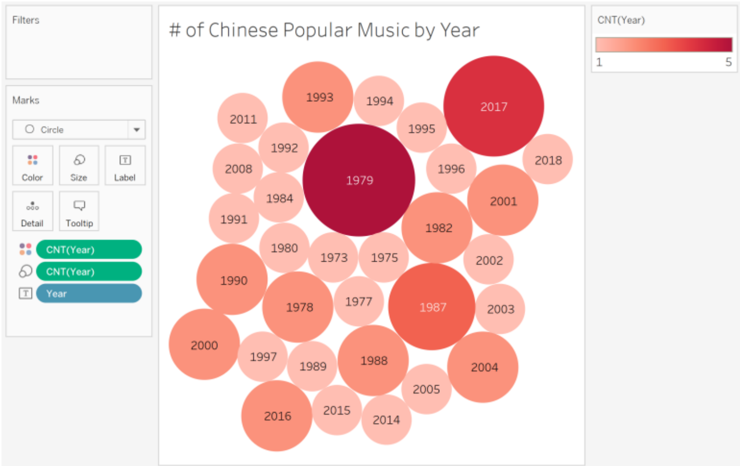

Figure 6 – No. of Chinese popular music by Year

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.

Commented [MK18]: While I know what a bubble plot is all about, you have to EXPLAIN because are your colours relevant? Or is it just the size?

Figure 7 – Word cloud of all English lyrics

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

Figure 9 – Word cloud of all Chinese lyrics

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'.

Commented [MK19]: This is a explanation that should accompany any of yourexploratory data analysus

Commented [MK20]: good

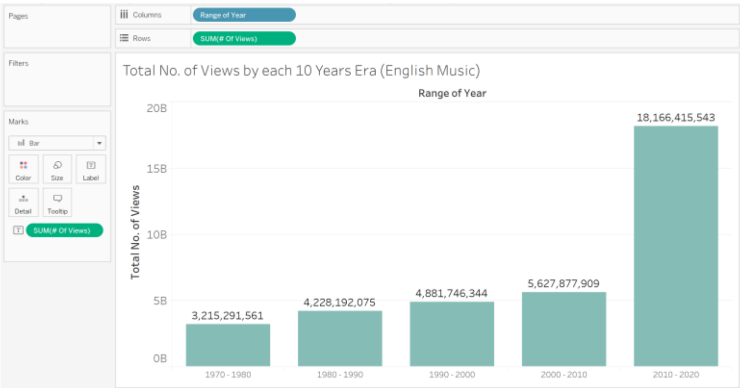

Figure 11 –No. of Views by Year Range (English Music)

Commented [MK21]: Using Tableau shows your versatility with software, but maybe not screenshot the whole thing? Just screenshot the graph. And if you are going to use Tableau, make your graphs professional looking, by adjusting the font size etc. I cant read anything; its too small

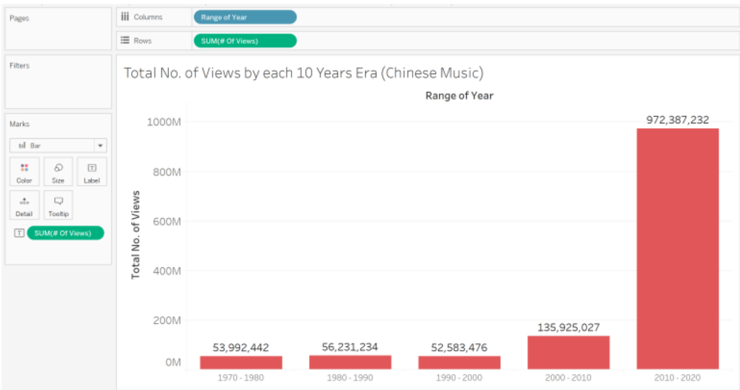

Figure 12 – No. of Views by Year Range (Chinese Music)

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).

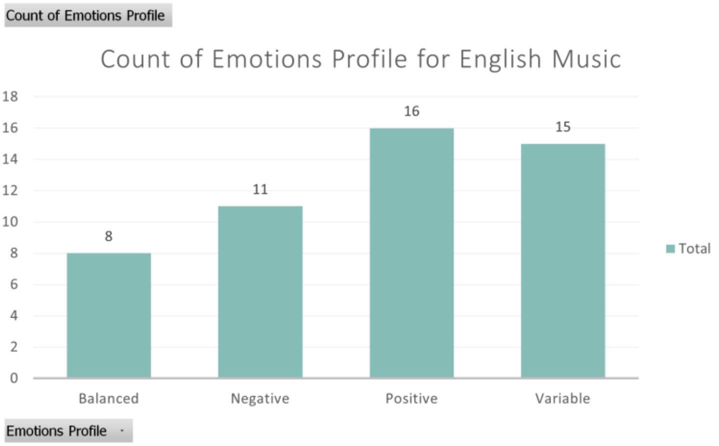

Figure 13 – Count of Emotions Profile (English Music)

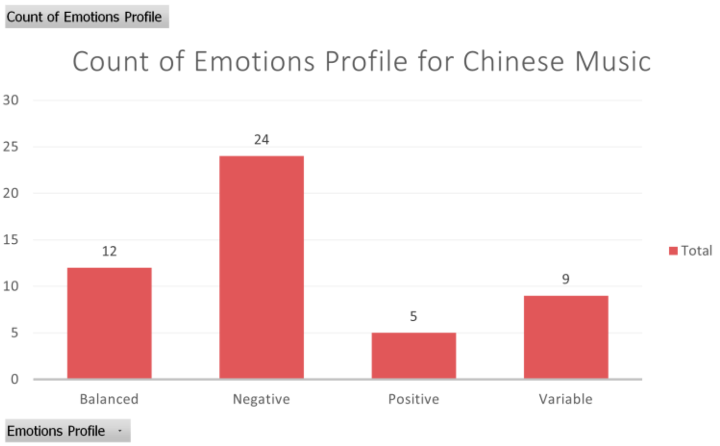

Figure 14 – Count of Emotions Profile (Chinese Music)

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.

Commented [MK22]: This is indeed interesting

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.

	Year	Music Title	Artist	Genre	Lyrics	Mood	Emotions Profile
0	1977	Stayin' Alive	Bee Gees	Pop	Well, you can tell by the way I use my walk\nI...	Energetic, Happy, Uplifting	Positive
1	1970	Layla	Derek and the Dominos	Rock	What'll you do when you get lonely\nAnd nobody...	Energic, Epic	Balanced
2	1978	Y.M.C.A	Village People	Rock	Young man, there's no need to feel down\nI sai...	Energetic, Happy, Sexy, Uplifting	Positive

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

	Year	Music Title	Artist	Genre	Lyrics	Mood	Emotions Profile
0	1978	夜來香	邓丽君	Rock	那南风吹来清凉\n那夜莺啼声细唱\n月下的花儿都入梦\n只有那夜来香\n吐露着芬芳\n我爱这...	Chill, Happy	Positive
1	1977	月亮代表我的心	邓丽君	Latin	你问我爱你有多深 我爱你有几分\n我的情也真 我的爱也真\n月亮代表我的心\n你问我爱你有多...	Calm, Chill, Romantic, Sad, Ethereal	Negative
2	1979	甜蜜蜜	邓丽君	Pop	甜蜜蜜 你笑得甜蜜蜜\n好象花儿开在春风里\n开在春风里\n在哪里 在哪里见过你\n你的笑容...	Calm, Chill, Romantic, Sad, Ethereal	Balanced

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

Commented [MK23]: Li Lin, it would make sense to draw this out with a single diagram simply? We don't need to see all your codes; that can be added in the Appendix.

I think bullet points or a simple flow diagram would suffice here

Text Pre-Processing for English Language

Data Cleaning:

1. Tranformed text to lowercase and duplicate it to a new column, cleaned_text.

```
# convert to lowercase into a new column
eng['cleaned_text'] = eng['Lyrics'].apply(lambda x: x.lower())
eng.head(3)
```

	Year	Music Title	Artist	Genre	Lyrics	Mood	Emotions Profile	cleaned_text
0	1977	Stayin' Alive	Bee Gees	Pop	Well, you can tell by the way I use my walk\nI...	Energetic, Happy, Uplifting	Positive	well, you can tell by the way i use my walk\ni...
1	1970	Layla	Derek and the Dominos	Rock	What'll you do when you get lonely\nAnd nobody...	Energic, Epic	Balanced	what'll you do when you get lonely\nand nobody...
2	1978	Y.M.C.A	Village People	Rock	Young man, there's no need to feel down\nI sai...	Energetic, Happy, Sexy, Uplifting	Positive	young man, there's no need to feel down\ni sai...

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.

```
contractions_re = re.compile('(%s)' % '|'.join(contractions_dict.keys()))
def expand_contractions(s, contractions_dict=contractions_dict):
     def replace(match):
        return contractions_dict[match.group(0)]
     return contractions_re.sub(replace, s)
eng['cleaned_text'] = [expand_contractions(i) for i in eng['cleaned_text']]
eng['cleaned_text'].head(3)
```

```
0    well, you can tell by the way i use my walk\ni...
1    what will you do when you get lonely\nand nobo...
2    young man, there is no need to feel down\ni sa...
Name: cleaned_text, dtype: object
```

Figure 18 – Sample after contractions

3. Replaced column, cleaned_text, values after removing punctuations and replacing new lines to space.

```
# removal of punctuations
def remove_punctuation(x):
    try:
        x = x.str.replace('[^\w\s]','')
    except:
        pass
    return x

eng['cleaned_text'] = eng['cleaned_text'].str.replace('[^\w\s]','')
eng['cleaned_text'] = eng.cleaned_text.replace('\n',' ',regex = True)
eng.head(3)
```

	Year	Music Title	Artist	Genre	Lyrics	Mood	Emotions Profile	cleaned_text
0	1977	Stayin' Alive	Bee Gees	Pop	Well, you can tell by the way I use my walk\nI...	Energetic, Happy, Uplifting	Positive	well you can tell by the way i use my walk im ...
1	1970	Layla	Derek and the Dominos	Rock	What'll you do when you get lonely\nAnd nobody...	Energic, Epic	Balanced	what will you do when you get lonely and nobod...
2	1978	Y.M.C.A	Village People	Rock	Young man, there's no need to feel down\nI sai...	Energetic, Happy, Sexy, Uplifting	Positive	young man there is no need to feel down i said...

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.

```
[('well', 'RB'), ('you', 'PRP'), ('can', 'MD'), ('tell', 'VB'), ('by', 'IN'), ('the', 'DT'), ('way', 'NN'), ('i', 'NN'),
('use', 'VBP'), ('my', 'PRP$'), ('walk', 'NN'), ('im', 'VBZ'), ('a', 'DT'), ('womans', 'JJ'), ('man', 'NN'), ('no', 'D
T'), ('time', 'NN'), ('to', 'TO'), ('talk', 'VB'), ('music', 'NN'), ('loud', 'NN'), ('and', 'CC'), ('women', 'NNS'), ('wa
rm', 'VBP'), ('ive', 'JJ'), ('been', 'VBN'), ('kicked', 'VBN'), ('around', 'IN'), ('since', 'IN'), ('i', 'NN'), ('was',
'VBD'), ('born', 'VBN'), ('and', 'CC'), ('now', 'RB'), ('it', 'PRP'), ('is', 'VBZ'), ('alright', 'JJ'), ('it', 'PRP'),
```

Figure 20 – Tokenization and POS tagging

```
['well', 'RB']
['tell', 'VB']
['way', 'NN']
['use', 'VBP']
['walk', 'NN']
['womans', 'JJ']
['man', 'NN']
['time', 'NN']
['talk', 'VB']
['music', 'NN']
['loud', 'NN']
```

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.

```
['well', 'tell', 'way', 'use', 'walk', 'woman', 'man', 'time', 'talk', 'music', 'loud', 'woman', 'warm', 'ive', 'kicked',
'around', 'since', 'wa', 'born', 'alright', 'okay', 'may', 'look', 'way', 'try', 'understand', 'new', 'york', 'time', 'ef
fect', 'man', 'whether', 'brother', 'whether', 'mother', 'stayin', 'alive', 'stayin', 'alive', 'feel', 'city', 'breakin',
'everybody', 'shakin', 'stayin', 'alive', 'stayin', 'alive', 'ha', 'ha', 'ha', 'stayin', 'alive', 'stayin', 'alive', 'h
a', 'ha', 'ha', 'stayin', 'alive', 'well', 'get', 'low', 'get', 'high', 'get', 'either', 'really', 'try', 'wing', 'heave
n', 'shoe', 'dancin', 'man', 'lose', 'know', 'alright', 'okay', 'ill', 'live', 'see', 'another', 'day', 'try', 'understan
d', 'new', 'york', 'time', 'effect', 'man', 'whether', 'brother', 'whether', 'mother', 'stayin', 'alive', 'stayin', 'aliv
e', 'feel', 'city', 'breakin', 'everybody', 'shakin', 'stayin', 'alive', 'stayin', 'alive', 'ha', 'ha', 'ha', 'stayin',
```

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.

```
chi['cleaned_text'] = chi['Lyrics'].apply(lambda x: x.lower())
chi.iloc[2:5]
```

	Year	Music Title	Artist	Genre	Lyrics	Mood	Emotions Profile	cleaned_text
2	1979	甜蜜蜜	邓丽君	Pop	甜蜜蜜 你笑得甜蜜蜜\n好象花儿开在春风里\n开在春风里\n在哪里 在哪里见过你\n你的笑容...	Calm, Chill, Romantic, Sad, Ethereal	Balanced	甜蜜蜜 你笑得甜蜜蜜\n好象花儿开在春风里\n开在春风里\n在哪里 在哪里见过你\n你的笑容...
3	1973	美酒加咖啡	邓丽君	Latin	美酒加咖啡\n我只要喝一杯\n想起了过去\n又喝了第二杯\n明知道爱情像流水\n管他去爱谁\...	Calm, Chill, Romantic, Sad, Ethereal	Negative	美酒加咖啡\n我只要喝一杯\n想起了过去\n又喝了第二杯\n明知道爱情像流水\n管他去爱谁\...
4	1975	再见我的爱人	邓丽君	Latin	Goodbye My Love\n我的爱人　再见\nGoodbye My Love\n相见不...	Calm, Chill, Romantic, Sad, Ethereal	Negative	goodbye my love\n我的爱人　再见\ngoodbye my love\n相见不...

Figure 23 – Example after case normalization

2. Replaced column, cleaned_text, values after removing punctuations and replacing new lines to space.

```
# removal of punctuations
def remove_punctuation(x):
    try:
        x = x.str.replace('[^\w\s]','')
    except:
        pass
    return x

chi['cleaned_text'] = chi['cleaned_text'].str.replace('[^\w\s]','')
chi['cleaned_text'] = chi.cleaned_text.replace('\n',' ',regex = True)
chi.head(3)
```

	Year	Music Title	Artist	Genre	Lyrics	Mood	Emotions Profile	cleaned_text
0	1978	夜來香	邓丽君	Rock	那南风吹来清凉\n那夜莺啼声细唱\n月下的花儿都入梦\n只有那夜来香\n吐露着芬芳\n我爱这...	Chill, Happy	Positive	那南风吹来清凉 那夜莺啼声细唱 月下的花儿都入梦 只有那夜来香 吐露着芬芳 我爱这夜色茫茫 ...
1	1977	月亮代表我的心	邓丽君	Latin	你问我爱你有多深 我爱你有几分\n我的情也真 我的爱也真\n月亮代表我的心\n你问我爱你有多...	Calm, Chill, Romantic, Sad, Ethereal	Negative	你问我爱你有多深 我爱你有几分 我的情也真 我的爱也真 月亮代表我的心 你问我爱你有多深 我...
2	1979	甜蜜蜜	邓丽君	Pop	甜蜜蜜 你笑得甜蜜蜜\n好象花儿开在春风里\n开在春风里\n在哪里 在哪里见过你\n你的笑容...	Calm, Chill, Romantic, Sad, Ethereal	Balanced	甜蜜蜜 你笑得甜蜜蜜 好象花儿开在春风里 开在春风里 在哪里 在哪里见过你 你的笑容这样熟悉...

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.

```
chi_lyrics = chi.cleaned_text
for eachLyrics in chi_lyrics:
    words = pseg.cut(eachLyrics)
    for w in words:
        print('%s %s' % (w.word, w.flag))
```

```
那 r
南风 nr
吹 v
来 v
清凉 a
  x
那 r
夜莺 n
啼声 v
細 a
唱 v
```

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.

	about	all	always	am	an	another	are	\
0	0.000000	0.039819	0.000000	0.000000	0.000000	0.093748	0.144527	
1	0.000000	0.061798	0.000000	0.000000	0.000000	0.000000	0.000000	
2	0.000000	0.097212	0.000000	0.000000	0.000000	0.000000	0.092589	
3	0.000000	0.149739	0.000000	0.000000	0.000000	0.000000	0.000000	
4	0.000000	0.135808	0.000000	0.000000	0.000000	0.000000	0.175875	
5	0.000000	0.041195	0.000000	0.000000	0.000000	0.000000	0.090040	
6	0.000000	0.000000	0.000000	0.000000	0.257624	0.000000	0.000000	
7	0.059538	0.062537	0.000000	0.000000	0.000000	0.000000	0.031036	
8	0.000000	0.129029	0.000000	0.000000	0.000000	0.000000	0.108857	
9	0.000000	0.104246	0.000000	0.000000	0.000000	0.000000	0.084134	
10	0.072168	0.036121	0.000000	0.000000	0.000000	0.250524	0.000000	
11	0.072168	0.036121	0.000000	0.000000	0.000000	0.250524	0.000000	
12	0.000000	0.133304	0.000000	0.000000	0.000000	0.000000	0.058181	
13	0.000000	0.089369	0.113809	0.000000	0.000000	0.000000	0.093078	
14	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.086284	
15	0.000000	0.000000	0.197370	0.330974	0.000000	0.000000	0.000000	
16	0.000000	0.042659	0.000000	0.000000	0.091981	0.000000	0.106023	
17	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	

Figure 26 – TF-IDF of English music lyrics

Commented [MK24]: So maybe somewhere at the start, you need to explain how the Chinese language is different from the English language. This is why I suggested a flow diagram or a comparison table of sorts

```
    children      city      cold  ...       黑暗        黑板        黑白       黑眼睛  \
0   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
1   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
2   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
3   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
4   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
5   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
6   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
7   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
8   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
9   0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
10  0.000000  0.000000  0.000000  ...  0.000000  0.003846  0.000000  0.000000
11  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
12  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
13  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
14  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
15  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
16  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
17  0.000000  0.000000  0.000000  ...  0.000000  0.000000  0.000000  0.000000
```

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.

Commented [MK25]: This TD_IDF was mentioned earlier; you should placed the definition there as well.

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 (IBM[1], 2020). However, this program is only available to analyze texts that are either English or French (IBM[2], 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 (MonkeyLearn[2], n.d.).

Commented [MK26]: I am just wondering here
(a)Would IBM WTA and Microsoft give the same results for the English sentiment score?
(b)Do you need to adjust your Chinse sentiment score given two different softwares? Or can we just assume that, should (a) be close enough (lets say 10%), we can accept and move on?

Commented [MK27]: Don't go into too much detail into the math but for the final report have some small subsections on the math and rationale behind these 2 methods. Why not random forest for example?

Commented [MK28]: What does this mean?

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

Commented [MK29]: good

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.

IBM[1] (2020). *About.* Retrieved August 14, 2021, from https://cloud.ibm.com/docs/tone-analyzer?topic=tone-analyzer-about.

IBM[2] (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.

MonkeyLearn[1] (n.d.). *Sentiment Analysis: A Definitive Guide.* Retrieved August 4, 2021, from https://monkeylearn.com/sentiment-analysis/.

MonkeyLean[2] (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.