Personalized Song Lyric Assistant

Edna C. Aldave, Mika Angelique C. Cortez, John Arvee R. Flores, Mideth B. Abisado, Larry A. Vea, and Dr. Yi-Hung Wu

The purpose of this study is to classify the emotion of a song by term frequency algorithm in your personalized song lyric collection using R language. The R language is widely used among data miners and statisticians for developing data analysis. Song lyrics are gathered from songlyrics.com only. Data was annotated by a psychology expert, cleaned the data and classified the emotion by word. Testing was done by comparing the annotated test data with the result from the application. Upon the completion of the study, identifying the emotion of a song lyrics with this type of method can be considered the easiest and its accuracy rate depends on the number of words in the dataset with emotions which is 40%. This research emphasizes the importance of emotion classification in Music Retrieval Information (MRI).

Index Terms – R programming language, Data Mining, Data Retrieval, Data Cleaning, Text Mining, Emotion classification

1. INTRODUCTION

A song is composed of hundreds of words which create its meaning. These collection of words is called song lyric which is a poem that express deep personal feelings written by artists. Music has been a part of the human culture and is said to play an important role in many social contexts. Emotion, a state of feeling, can be divided into two types which is related to music. These are perceived emotion and felt emotion. Perceived emotion refers to what emotion the listener identifies what is invoked by the music while felt emotion is the emotion felt while listening to the music. [11]

In the field of Music Information Retrieval, the classification of music genre and emotion has become the main focus of recent studies. Music Emotion Classification (MEC) lies in between these two challenging areas of research since the number of music recordings is rapidly increasing. [11]

In this paper, the researchers aimed to create a model classifying music emotion using lyric features of English songs. The researchers used Paul Ekman's six basic emotions as emotion model. The songs used as data were classified into the following genres: country, hip-hop, pop, rock and R&B [11]. The music lyrics were annotated by an expert and were also extracted by Bag-of-Words feature representation. The emotion of the lyric feature were processed by counting the term frequency of each word in a song. The main objective of this study is to classify the emotion of the song in the collection and to create a decision support feature by recommending related songs based on the emotion of the selected song. This research will help the user in providing relevant songs related to the user's collection.

The researchers used R programming language as a tool for developing the program. The experimental data used were limited from www.songlyrics.com and www.wikipedia.org. The web application can recommend related songs based on classifying the emotion of the song.

1.1 THEORETICAL FRAMEWORK

Term Frequency (TF) measures the frequency of a term in a document. The term frequency is usually divided by the document length as a way of normalization since the documents differ in length [21]

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

Tokenization is the process of segmenting text into phrases, words, symbols and other

meaningful elements into terms or tokens in lexical analysis. For further processing, the list of tokens becomes the input in parsing or text mining. A simple tokenizer can break the string up into terms wherever it encounters whitespace or punctuation. [27]

Text Mining Framework

Since texts are unstructured collections of words, text mining analysis was influenced to consist numerous number of challenging steps. For the first step, texts should be imported in the computer environment which is in this case R. After the texts has been organized in a repository, the next step is pre-processing and tidying up the texts to achieve a convenient representation for future analysis. Text reformatting (e.g., whitespace removal), stopword removal, or stemming procedures might be included in this step. The third step includes the transformation of the preprocessed texts into structured formats to be actually computed in. This basically implies the creation of the term document matrix which is the most common format to represent text in a computation. So, the analyst can work and compute on texts using clustering or classification methods which are standard techniques from statistics and data mining. A text mining framework should accomplish the following: offer functionality for managing text documents, abstract the process of document manipulation and ease the usage of heterogeneous text formats. But there is a necessity to a database-like conceptual entity in holding and managing text documents in a generic way which we call a text document collection or corpus.

Text mining usually involves doing computations on texts to gain interesting information. The most common approach is to create term-document matrix to hold the frequencies of distinct terms for each document. Computing directly on character sequences as is done by string kernel methods is another approach. Basically, the infrastructure and framework supplied by tm intends to implement the conceptual framework discussed.

1.2 CONCEPTUAL FRAMEWORK

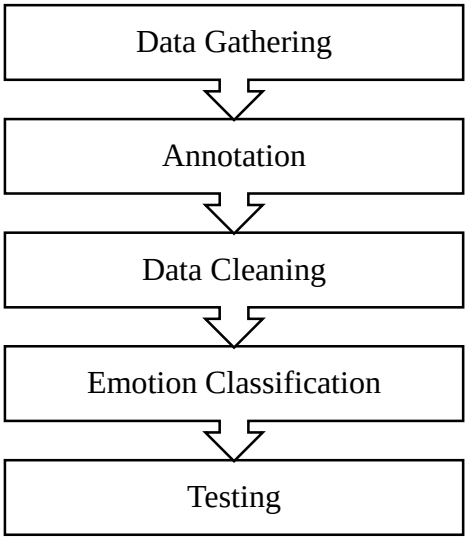

Figure 1.1 Conceptual Framework

The conceptualized framework explains the flow of this study to answer the research questions. The data were collected from the websites: www.songlyrics.com and www.wikipedia.org. Songs collected were annotated by an expert based on Paul Ekman's six basic emotions. The researchers used the tm package in R to create a corpus of the whole song lyrics and then clean the data by removing the stopwords, unnecessary elements and spaces. A dataset of words with its corresponding emotion were obtained from the sentiment package. Term frequency was the algorithm used in the corpus of the song lyrics which is based on the dataset. Testing was done by comparing the results from the application and the expert's annotation on the emotion of the song.

2. REVIEW OF RELATED LITERATURE

R Programming

In this study, the programming language that will be used by the proponents is R. R is a programming language and software environment that is used for statistical computing

and graphics. It is originally created and distributed by Bell Labs. [1] The R environment is an integrated suite of software facilities for data manipulation, calculation, and graphical display. It has an effective data handling and storage facility; operators for calculations on arrays and in particular matrices; collection of tools for data handling; graphical analysis for data analysis, programming language that includes conditional loops, user-defined recursive functions, input and output facilities. R has different packages that can be used for different functions which is available in CRAN.

R Packages

Packages that deal in designing the interface of the web application. **Shiny** is a web application framework for R which makes analyzed data into an interactive web application. For further modifying of the interface, package like **Shinydashboard** is used to extend the provided features of Shiny. To provide a message box for the interface, **SvDialogs** is used. **Rvest** provides functions which deals with data retrieval from web pages. With the help of Selector Gadget which is a CSS selector extension for a web browser, it will make the web scrapping easier. **Downloader** package provides the function on downloading files over HTTPS. Manipulating the data makes **Stringr** package to be useful specifically in strings. It simplifies the implementation of basic operations on the strings by providing a clean and modern interface. **Data.table** offers a simple but flexible syntax in manipulating the data and placing it into tables which can provide large data. This package is also used in different fields like finance and economics. **NLP** is a package dependency of **TM** which is used for text mining. It provides methods in data import, corpus handling, preprocessing, metadata management and term-document matrices. **Wordcloud** displays different ways on presenting a word cloud. Most applications uses wordcloud package for presenting the frequency of words using their data. **Sentiment** is an R package with tools for sentiment analysis including bayesian classifiers for positivity/negativity and emotion classification.

Emotion of Music

Emotion is a common intuitive reflex state of mind inferring from one's circumstances, mood, or connections with others. Identifying the emotion of music has been a study for researchers now. It may be interesting to determine the emotion of a music for it may be helpful to users. Music is meaningless without emotions, emotion is the essence of music [12]. "In a nutshell, the value for a listener of expressiveness in music, understood as its ready hearability as personal expression, is the value of confronting images of human experience…images which are in a sense woven out of the substance of music, as substance in which one…becomes the music while listening, and so participates in the mental life embodied in the music" [13]. There are different sets of emotions to be considered in classifying the emotion of a song like the Thayer model and Paul Ekman model. In the study of Isidro and Noblejas, [11] the emotion model that they used in the Paul Ekman model which consist of 6 basic emotions: Happiness, Sadness, Anger, Fear, Disgust, and Surprise [11]. An emotion may be happy if it gives contentment, satisfaction, pleasure, joy. Sadness if it is about disadvantage, loss, and helplessness. Furthermore, sorrow, grief, melancholy, misery, temporary lowering of mood. It's Anger if its feeling may range from minor irritation to intense rage. Disgust if it's unclean, inedible, infectious, offending. Feeling of Fear when it's emotional response to threats and danger; related to escape or avoidance. And lastly, Surprise which may result of experiencing an unexpected relevant event according to the theory of Paul Ekman, a pioneer psychologist of emotions. [14]

Emotion Classification Model

Classifying emotions in music can be done using different models. Emotion classification is distinguishing the emotion or mood from another is a challenge in the field of emotion research and affective science. [15] There are various models

used for emotion classification. Some of these include: circumplex model, vector model, PANA (positive activation – negative activation) model, plutchik's model, PAD emotional state model [16]. While in the study of Isidro and Noblejas, SVM, Naïve Bayes, and k-NN models were used. In this study, emotion is being classified through words or lyrics since the study is about identifying the emotion of a song through its lyrics.

Bag-of-words is usually used in reading texts or document classifications wherein the occurrence of each word is used as a feature in training a classifier. It is commonly used for analyzing text documents by counting the occurrences of words or frequency but not considering the grammar of sentence or it ignores the semantics of the words. [6]

Term frequency, Inverse Document Frequency (tf-idf) is a numerical statistic used to determine how important a document is in a corpus. [20] It is used frequently for weighting words in text mining and information retrieval. It is in the term of scoring of words to see its importance and how they frequently appear in the document. [18] Tf-idf is commonly used to filter out stop-words so it is relevant for text summarization and text classification. It works as if the word frequently appears then the word is important so the weight of it must be high but if the word appears only in different documents then the weight must be low therefore it's not a unique identifier. [18]

Term Frequency (tf) is also a numerical statistic used for text mining and text classification. There is somehow a difference in term frequency (tf) and term frequency-inverse document frequency (tf-idf). Term frequency is also for getting the weight of words in a document. In an article about term frequency, it is said that, " In this view of a document, known in the literature as the bag-of-words model , the exact ordering of the terms in a document is ignored but the number of occurrences of each term is material (in contrast to Boolean retrieval). We only retain information on the number of occurrences of each term. Thus, the document example, ``Mary is quicker than John" is, in this view, identical to the document ``John is quicker than Mary". Nevertheless, it seems intuitive that two documents with similar bag of words representations are similar in content."[21] Technically, if term frequency is not paired with idf (inverse document frequency) then it would just be getting the weight of the terms in the document. The simplest way to use term frequency is to count the number of the uses of each key word in a page. [22] In this study, the package that is used for term frequency is the tm (text mining) package in R.

Machine learning algorithm is exploring the study of algorithms that can be learned from making predictions on data. There are three (3) types of Machine Algorithm namely: Supervised Learning Algorithm wherein there is a target or outcome variable hat is to be predicted from a set of variables; Unsupervised Learning, there is no target or outcome to predict; while Reinforcement Learning, to make the machine have specific decision. Some common machine learning algorithm used are: Linear Regression, Logistic Regression, Decision Tree, SVM, Naïve Bayes, KNN, etc. [23]. These machine learning algorithms can be used in R programming.

Text Mining is the analysis of data in natural language text. The application text mining is called text analytics. [24] Text mining involves the process of making a structure of an input like parsing. Tasks in text mining include: text categorization, text clustering, entity extraction, document summarization. [25] Text analytics on the other hand, it is in a form of software that helps unstructured data of words and phrases to numerical values in which it can be linked to structured data in databases and be analyzed through data mining. [24]

WordNet is technically a lexical database wherein it is a combination of dictionary and thesaurus in the English language. It is used for text analysis and artificial intelligence applications and softwares. [26]

Data Mining is the process of analyzing data and making it into useful information by summarizing in different perspective. [4] Parameters in data mining include: Association, where patterns are looked for and one event is connected to another; Sequence or Path analysis, where patterns are looked for and one event leads to another later event; Classification, new patterns are looked for; Clustering is when facts that are not known is documented visually; lastly is Forecasting, where patterns in data are discovered that can lead to reasonable predictions in the future. [5]

3. METHODOLOGY

Data Gathering

The data which the proponents gathered came from two websites: www.songlyrics.com and www.wikipedia.org. The song lyrics to be used was 20 English songs which were randomly picked from the website. These songs were used as test data while for the dataset, the proponents used the corpus from the sentiment package and also from the NRC Emotion Lexicon collection of S. Mohammad. Songs added in the collection by the user can also be predicted using the emotion classification.

Annotation

After trimming the song lyrics into chorus part and saving it into text files, the data were sent to an expert which is a graduate of BS in Psychology in Centro Escolar University. The researchers explained to the expert the instruction of the annotation. The annotation was done through classifying the emotion of each line of the chorus part using Paul Ekman's six basic emotion model.

The annotation were completed by a week. After collecting the annotated data, the researchers tallied the emotion of each chorus and the emotion which was given to the majority of the line was chosen as the emotion of the song.

Data Cleaning

After collecting the annotated songs, song stanzas were placed as variables to prepare for the data cleaning process. Below is an example of a chorus part used in the testing of the classifier. Using bag-of-word approach, lyric stanzas were sliced into unigrams also to determine the words which carry emotions based from the dataset.

```
> x <- strsplit(song_test_data[10], " ")
> x
[[1]]
 [1] "Get"        "high"       "baby"       "roll"       "one,"
     "cloud"      "nine"       "'bout"
 [9] "to"         "go"         "up"         "Lovin'"     "the"
     "feelin'"    "the"        "turbulence"
[17] "get"        "when"       "we"         "turn"       "up"
     "when"       "we"         "land"
[25] "we"         "can"        "roll"       "out"        "show"
     "you"        "somethin'"  "you"
[33] "ain't"      "know"       "about"      "Tonight"    "we"
     "be"         "takin'"     "off"
[41] "flight"     "with"       "a"          "camera"     "to"
     "show"       "out"
```

Figure 3.1 Song Test Data

Through tm() package, the first thing which was removed from the corpus were the numbers. Aside from numbers, punctuations and whitespaces were also removed. The words were also transformed into lowercases to prevent the frequency of same word but different cases. Stopwords were also removed since they are common to words in a language and do not contain important significance.

```
> corpus_read  <- Corpus(VectorSource(song_test_data[10]), readerControl
= list(reader=readPlain))
> corpus <- tm_map(corpus_read, content_transformer(tolower))
> corpus <- tm_map(corpus, removePunctuation)
> corpus <- tm_map(corpus, stripWhitespace)
> corpus <- tm_map(corpus, removeNumbers)
> corpus <- tm_map(corpus, removeWords, c(stopwords("english")))
> dtm <- TermDocumentMatrix(corpus)
> test_song_edited <- dtm$dimnames$Terms
> test_song_edited
 [1] "aint"       "baby"       "bout"       "camera"
 [5] "can"        "cloud"      "feelin"     "flight"
 [9] "get"        "high"       "know"       "land"
[13] "lovin"      "nine"       "one"        "roll"
[17] "show"       "somethin"   "takin"      "tonight"
[21] "turbulence" "turn"
```

Figure 3.2 Song Test Data cont.

Emotion Classification

Sentiment package is an R package with tools for sentiment analysis including bayesian classifiers for positivity/negativity and emotion classification. Classify_emotion() function was used in this study to predict the emotion of a song.

```
> song_emotion <- classify_emotion(test_song_edited, algorithm = "bayes",
 prior = 1.0)
> best_fit = song_emotion[,7]
> data <- cbind(song_emotion,test_song_edited)
> emotion_table <- sort(table(best_fit), decreasing = TRUE)
> final_emotion <- names(emotion_table[1])
> emotion_table
best_fit
  anger     joy    fear disgust sadness
      5       3       2       1       1
> final_emotion
[1] "anger"
```

Figure 3.3 Emotion classification of song test data

After cleaning the data, words were subjected for emotion classification, each word weighs an emotion based on the dataset of the sentiment package and on NRC Emotion Lexicon by S. Mohammand. [2]

```
> freq
    get    roll    show    aint    baby    bout  camera     can   cloud  feelin
      2       2       2       1       1       1       1       1       1       1
 flight    high    know    land   lovin    nine     one somethn    talkn   tonight
      1       1       1       1       1       1       1       1       1       1
turbulence    turn
      1       1
```

Figure 3.4 Term Frequency

Words were matched from the dataset and were counted by the emotion each word is carrying. The emotion model used for this study is based on Paul Ekman's six basic emotion. Counting all the words pertaining to each emotion was also done and the emotion which contains most number of words from the song was chosen as the emotion of the song. Looking at appendix B will provide the whole computation for the example below.

Testing

In evaluation using the emotion classifier, accuracy was measured. Accuracy is defined as the percentage of correctly identified instances by the model. This study rated the performance of the emotion classifier based on its accuracy.

To further determine the accuracy of the used emotion classifier, 20 songs as test data were used for testing. 20 English songs were extracted into bag-of-words and were subjected for annotation of the emotion of the song. Test data was used in predicting the emotion of each song to determine if the model was able to correctly identify the emotion of the song compared to the annotator's answers.

4. RESULTS AND DISCUSSION

To determine the accuracy, testing was done using twenty songs (randomly selected) annotated by the expert. This dataset was used as test data to determine whether the emotion classifier could be applied to unknown data.

Out of twenty songs predicted by the emotion classifier, only eight songs were predicted correctly having 40% accuracy rate of classifying the emotion of the song.

The table below shows the number of correct predictions of the emotion classifier based from the annotated answers. The emotion classifier identified songs from top five different song genres (four songs per genre).

NO. OF CORRECT PREDICTIONS (GENRE)
(Four songs per genre)

GENRE	Country	Hip-Hop	Pop	Rock	R & B
SCORES	1	2	2	3	0

Table 1. Number of Correct Predictions (Genre)

The relationship of the classifier and the genres were also observed. It can be concluded that three out of four songs with Hip-hop genre from the test data were predicted correctly while songs with R&B genre yields the lowest accuracy in predicting the emotion of the song.

NO. OF CORRECT PREDICTIONS (EMOTION)

EMOTION	H	SA	A	F	D	SU
SCORES	5	1	2	0	0	0

Table 2. Number of Correct Predictions (Emotion)

Looking closely on the prediction results of the emotion classifier, it can be said that predicting the emotion label "Happiness" was the easiest since most of the correct predictions were under the "Happiness" emotion label. However, predicting the emotion based on lyric features only may be difficult. The researchers concluded that the data set may affect the prediction results

since majority of the lines from the data set used were annotated as “Happiness”. Providing a better data set could lead to better accuracy rate in classifying the emotion.

For the process of identifying the related song based on emotion, the researchers created a function that retrieves 5-10 randomly selected songs from the website. These songs undergo the process of data cleaning and classifying the emotion same as how training and test data were processed in getting their emotion. After classifying randomly selected songs, songs with the same emotion with the selected song will be picked as the related song.

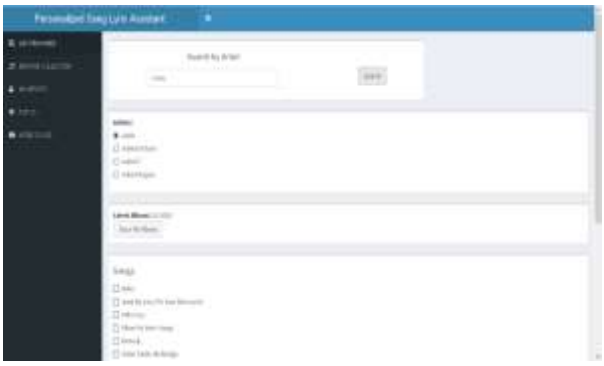

Figure 4.01. Get from Web user interface

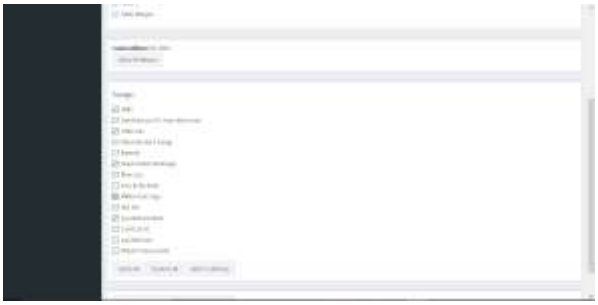

Figure 4.02. Get from Web user interface cont.

Figure 4.03. Get from Web user interface cont.

Figure 4.01 shows the get from web user interface. In this tab, the user can search the name of the artist that he/she wants. It will show the list of artist where the user can chose their desired artist. The application will show the list of the songs of the latest album of the artist. The user has an option to show all albums of the artist. They can select the songs that they want to add in their collection which is shown in figure 4.02. As for the show all albums option, all albums of the artist will be shown where the user can choose which album he/she is interested in, in order to add their desired songs in their collection that is shown in figure 4.03.

Figure 4.04. Browse Collection user interface

Figure 4.05. Browse Collection user interface

The data retrieved from the web is stored in text files and are shown and organized in the browse collection tab. Figure 4.04 shows the collection of the user based from what he/she added to their collection. The data are sorted by columns which are: Songs, Artist, Album, and Genre. Once the user clicks a row item in the table, the emotion of the song will be displayed and also the lyrics of the song as shown in Figure 4.05. Furthermore, related song will be displayed if ever the program has retrieved a random song which is classified as

the same emotion. However, a button is placed beside to gather more songs in order to retrieve a related song based on its emotion. The user also has an option to delete songs from their collection by selecting the items and clicking the "Delete" button. Searching a song through the user's library is also applicable.

Figure 4.06. Top 10 user interface

Figure 4.07. Top 10 user interface (by Genre)

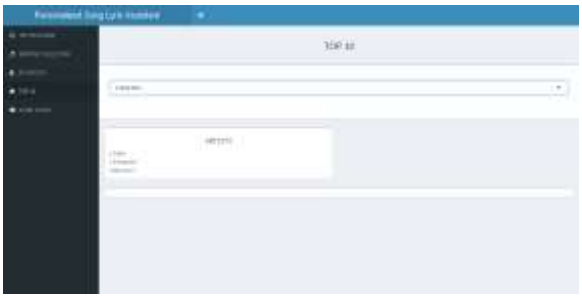

Figure 4.08. Top 10 user interface (from collection)

In the Top 10 tab of the application, it shows the top 10 songs and artists from the website. The user has an option to choose whether the top 10 artist and song will be coming from the web or his/her collection. In figure 4.06, it shows the top 10 songs and artist from the website (www.songlyrics.com) from all genre. The user also has an option to see the top 10 songs and artists based on genre chosen in the genre option above which is shown in figure 4.07. While in figure 4.08, it shows that the user can choose the "from collection" option and this will show a different result. It only shows the top 3 artists based on the collection, it will be dependent on the number of songs per artist. This where data analysis comes in.

Figure 4.09. Word Cloud user interface

This figure shows the visualization part of the study. The proponents used the package wordcloud() in order to show the word cloud of the most used words per artist. The two word cloud visualization explains that there can be two different artists that can be compared based on the words that they use in their songs. This word cloud displays at the maximum of 100 most used words in their songs. These words are weighted and analyzed and is shown in a form of word cloud as a result.

5. CONCLUSION AND RECOMMENDATION

Conclusion

In this paper, the researchers created a program entitled Personalized Song Lyric Assistant which was produced using R programming language. Most of the words used as dataset were also annotated with the emotion "happiness". Based on the result of the research, the emotion induced by English songs is, generally, "happiness". For the testing process of the model, results showed that songs from genres Hip-hop can be clearly classified by the emotion classifier. This might be due to the reason of having a dataset composed,

generally, with happiness emotion. Classifying emotions like fear, disgust and surprise might have less accuracy than happiness and sadness since the dataset used was composed of fewer words carrying those emotions.

Based from the performance measures used in classifying the model, results showed that by using term frequency using bag-of-words approach to represent lyric features were able to achieve at least 40% accuracy rate.

For the results of getting related songs, using unknown data, the model was able to classify their emotions through processing the whole lyrics just the same as how the test data were processed. Unknown data undergo also in data cleaning, representing it in unigrams and counted the frequency of the terms in order to classify the emotion of the song.

Creating a program which provides sentiment analysis was made easier with the help of R language and RStudio as a tool in developing the application. Different packages were used in order to build the whole application. However, a problem was encountered about the program reaching the timeout on getting the data from web due to slow connection of internet or because of the default length of time in retrieving a data in R language. Aside from data retrieval, there are also packages which are not compatible with each other.

Recommendation

For improvements of further studies using R language, the researchers suggest to have more research on every package that will be used since compatibility is a major reason on considering of using it. As for data retrieval, providing a better algorithm will solve the problem on causing the application on reaching timeout on getting data from a website or page.

Based on the results of the study, some recommendations are made to improve further studies related to sentiment analysis, specifically in emotion classification:

• A larger dataset of words with emotions to avoid unknown words and achieve higher accuracy
• Using Tf-idf as a bag-of-words approach may yield better accuracy rate in classifying the emotion of the song.
• Use a model in classifying the emotions

REFERENCES

[1] Find and Analyze Data (2015). Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-r.html

[2] Mohammad, S., Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon

[3] Howard, S., Silla Jr., C. N., & Johnson, C. G. (2008). Automatic Lyrics-based Music Genre. 1-4.

[4] Data Mining: What Is Data Mining (2011). Retrieved from: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[5] Rouse, M. (2012). Data Mining. Tech Target. Retrieved from: http://searchsqlserver.techtarget.com/definition/data-mining

[6] Tran, K. (2014). Alternatives for Analyzing Text Documents. Retrieved from https://www.quora.com/What-are-the-alternatives-to-bag-of-words-for-analyzing-text-documents

[7] Dannenberg, R., Hwang, E., Han, B., Rho, S. (2009). Smers: Music Emotion Recognition Using Support Vector Regression

[8] Malheiro, R., Panda, R., Gomes, P., Paiva, R. (2010). Music Emotion Recognition from Lyrics: A Comparative Study

[9] Strapparava, C., Mihalcea, R. (2011). Lyrics, Music, and Emotions

[10] R: Documentation. Retrieved from https://www.r-project.org/other-docs.html

[11] Isidro, D., Noblejas, K. (2015). Emotion Detection of Filipino Music

[12] Omar (2011) How Important Is Emotion in Music? Retrieved from: https://ph.answers.yahoo.com/question/index?qid=20110112124456AAZofQW

[13] Levinson, J. (2011) Music, Art, and Metaphysics, ch. 14, PP. 336-375. See especially Part I, pp. 337-357.

[14] Tsorli (2011) 6 Basic Emotions Retrieved from: https://www.behance.net/gallery/930168/6-Basic-Emotions

[15] Colombetti, Giovanna (August 2009). "From affect programs to dynamical discrete emotions". Philosophical Psychology 22 (4): 407–425.doi:10.1080/09515080903153600.

[16] Rubin, D. C.; Talerico, J.M. (2009). "A comparison of dimensional models of emotion". Memory 17: 802–808. doi:10.1080/09658210903130764.

[17] Omar (2011) How Important Is Emotion in Music? Retrieved from: https://ph.answers.yahoo.com/question/index?qid=20110112124456AAZofQW

[18] Loria, S. (2014). Tutorial: Finding Important Words in Text Using TF-IDF Retrieved from: http://stevenloria.com/finding-important-words-in-a-document-using-tf-idf/

[19] Tsorli (2011). 6 Basic Emotions Retrieved from: https://www.behance.net/gallery/930168/6-Basic-Emotions

[20] Rajaraman, A.; Ullman, J. D. (2011). "Data Mining". Mining of Massive Datasets (PDF). pp. 1–17.doi:10.1017/CBO9781139058452.002.ISBN 9781139058452.

[21] Term Frequency and Weighting (2008). Retrieved from: http://nlp.stanford.edu/IR-book/html/htmledition/term-frequency-and-weighting-1.html

[22] Enge, E. (2015). Using Term Frequency Analysis to Measure Your Content Quality. Retrieved from: https://moz.com/blog/using-term-frequency-analysis-to-measure-content-quality

[23] Ray, S. (2015). Essentials of Machine Learning Algorithms (with Python and R Codes). Retrieved from: http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/

[24] Rouse, M. (2010). Text Mining. Tech Target. Retrieved from: http://searchbusinessanalytics.techtarget.com/definition/text-mining

[25] Wayback Machine (2009). Retrieved from https://en.wikipedia.org/wiki/Text_mining

[26] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. (1990). WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.

[27] Huang, C., Simon, P., Hsieh, S., & Prevot, L. (2007) Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Word break Identification.