Leveraging social media in the music industry

An investigation of Twitter analytics as an input for the prediction of song performance in music charts

by

Hoang Oanh Le

2019

Masters Thesis

Masters Thesis submitted in accordance with the rules of
TIAS Business School
in partial fulfillment of the requirements for the degree of
International MSc in Business Administration
Specialization Business Analyics

TIAS

STATEMENT OF AUTHENTICITY

I have read the TIAS Regulations relating to plagiarism and certify that this project is all my own work and does not contain any unacknowledged work from any other sources

I confirm that the Word Count as per the TIAS Regulations is 13950 words

Signed : Hoang Oanh Le

Date: 30/08/2019

Name: Hoang Oanh Le
Title: Leveraging social media in the music industry: An investigation of Twitter analytics as an input for prediction of song performance in music charts
Student Number: u462064

Abstract

Social media is seen as a platform where people freely express their opinions about any matter, thus, generating a massive amount of user-generated content. Twitter undoubtedly has held its firm position among all social networking sites with an exponential number of users every year. Many studies were carried out by investigating the power of Twitter data in health care industry, politics, sports, and music industry. Over the last five years, the music industry has experienced a shift in the way people listen to music since the introduction of online streaming music. Music lovers are prone to interact with their favorite songs and artists through social media, which provides enormous troves of insight not on just individual song and artists but also on how music consumers perceive any song. Therefore, many kinds of research have been carried out to investigate the impact of Twitter on forecasting songs revenue. However, there are only two studies that aimed to explore the predictive power of Twitter to song performance. This paper shed some light on this little-recognized topic by evaluating Twitter data in forecasting song popularity, which is demonstrated via the Billboard Top 100 chart.

The results indicated that while Twitter data can be utilized as a predictor of song popularity, incorporating Twitter and Billboard information (number of weeks the songs presented in the chart) enhance chart prediction than sole Twitter data. Findings of this study are beneficial to the music industry to discover song performance by real-live update trends on social media in order to propose an appropriate strategy for hit and non-hit songs.

Keywords: user-generated-content, text mining, Python, machine learning algorithms, supervised classification

Preface

Having a particular interest in social media and spending hours a day catching up with the latest trends in the world through Facebook, Instagram, and Twitter, I am more than familiar of how social media influence our life. Also, I am a big music lover who share listening behaviors frequently on the internet with like-minded people and let music unconsciously be an integral part of my life. Hence, combining all these interests with my passion for data, I chose to explore the power of Twitter in music industry, particularly in predicting song performance.

This thesis is dedicated to my family and my friend - thank you for your endless support and encouragement during my master study period. And another important person who has been through ups and downs with me, patiently guided me through difficulty in machine learning knowledge even though we are a half-world apart from each other – my boyfriend. I would also like to thank TIAS School for Business and Society and Professor Zahid Hussain for guidance and support provided in this journey. This accomplishment will never be possible without them.

Hoang Oanh Le
30th August 2019

Table of Contents

List of Figures

List of Tables

List of Twitter Jargons

Twitter has a set of primary features used in this study are described as follow:

- A tweet: Twitter users often post periodic status, which is called tweet, which contains less than 140 characters. These tweets can vary from a broad range of topics which typically consist of personal information of users, the news, and links to content they share. (Help.twitter.com, n.d.)
- A hashtag is indicated by a word preceding with # this symbol. It is used to index keywords or topics on Twitter, which allows users to follow interesting topics (Help.twitter.com, n.d.)
- A mention is a tweet than contains another username anywhere in the body of the tweet by using symbol @ to notify those users that they are mentioned in the tweet. (Help.twitter.com, n.d.)

Chapter I: INTRODUCTION

1.1 Background information

Nowadays, in our booming era, the freedom of expression on the internet has been simplified than ever. The widespread adoption of social media has empowered users to freely express themselves and engage in far-flung conversations with people from all over the world. Social media, therefore, presents academic researches with information from a wide range of topics within a naturally occurring setting. Given the significance of social media, understanding user-generated content provides a rich source for management and strategists in various industries. It is no exception to the billion music industry, which has experienced an exponential growth thanks to the introduction of streaming music over the last ten years. The fast growth of technology has transformed traditional music platforms into online music services where everyone is brought closer together. Social media platforms are where music audiences naturally congregate, building their own communities, and sharing experiences of songs and artists that subsequently is a catalyst to encourage users' engagement to the music industry.

Among top common social networking sites, Twitter remains as the best indicator of the broader pulse of the world and overview picture of what is happening within (Hutchinson, 2016). It has become an outlet for people to chatting about music. According to Brandwatch's Twitter Landscape Report 2013, music is the third-most talked topic on Twitter, following TV and sports (Franklin, 2013). Moreover, in fact, as of June 2019, 6 out of the top 10 most – followed Twitter accounts are all musicians (Cement, 2019). Additionally, music audiences tend to share what they are listening and their feeling of a song on social media. In a new study by Billboard (2018), it is found that 29% of all social media users are likely to share songs, albums, or playlists from streaming services. Given a considerable amount of music-related users' behaviors data, understanding Twitter is a potential source for the music industry in various perspectives. With all of this connection between Twitter and the music industry, the initial question that comes to mind of the author is "to what extent Twitter data can predict song rankings?". The observation of the author also inspires this subject as a big music fan that popular songs are likely to be shared and discussed on Twitter than less well-known ones. This study will hence focus on metrics

retrieved from Twitter with the main goal is to explore the contribution of online chatter in defining song rankings.

1.2. Research Motivation

1.2.1 Literature Perspective

The integration of social media and music industry is not a new concept but rather has been exploited in many previous researches. Concerning researches that use Twitter as the primary data source, a study of Schedl and Tkalcˇicˇ (2014) found that classical music listener does not use social media platform like Last.fm and Twitter to post their listening behaviors as frequently as listeners of other music types. The listening behaviors of music audiences on Twitter also contribute to building a corpus for music recommendations (Zangerle, Gassler, and Specht, 2012). In another research of Zangerle, Pichl, Gassler, and Specht (2014), tweets that contain #nowplaying were exploited to extract listening behaviors from Twitter users. Hashtag #nowplaying can be considered as a powerful tool to link Twitter with the music industry because people are likely to share what they are listening on Twitter via #nowplaying. There are two previous researches by Kim (2014) and Zangerle, Pichl, Hupfauf, and Specht (2016), which also leveraged the power of Twitter in predicting music charts by exploring #nowplaying tweets. Even though Twitter data was found as a predictor for accurate song rankings forecast, these studies were limited in addressing volume-related variables only (these studies will be further discussed in detail in 2.3: Song rankings). None of them have explored the power of sentiment-related variables. Hence, the undertaken dissertation found a research gap that could be investigating the predictive power of Twitter by addressing both volume-related and sentiment-related variables.

1.2.2 Author Perspective

While literature perspective provides the author the academic-based goal for investigation, another key fact to take into account is that the author is a typical music lover, live and breathe with music. The social media platform is where music lovers facilitate the conversation with others to share their expression towards any song. This has triggered the author to uncover the relationship between songs popularity and online chatter. To that end, it is both a matter of having a personal interest in the

research field while helping the music industry to leverage the social media platform for management purposes.

1.3. Research Objectives

The primary goal of this research is to add values to management and marketers in the music industry in managing online chatter, creating better marketing strategy by evaluating the contribution of Twitter feeds towards defining the song rankings performance. Additionally, the result is expected to shed light on the behaviors of music listeners in social media. Regarding a limited number of studies leveraging Twitter data in the music industry, the research findings are expected to contribute to mining Twitter data area, and the predictive model can be applicable for post-study in the music industry as social media has increasingly been highly involved in our daily life and music is an inseparable part of it.

1.4. Research Questions

How useful is Twitter data as the mean of input for the prediction of the song ranking chart?

The following sub-questions have been formulated based on main research questions:

- What is the relationship between each independent variable and song ranking?
- Whether Twitter data solely can predict the song ranking?

1.5. Research Expectations

Through exploratory data analysis of a real-life context, the expectation is that this study provides a rationale and scientific answers to questions related to the correlation of online chatter and song rankings. It is notably mentioned that the lack of time and data created an opportunity to see whether Twitter data is possible to predict song ranking within a short period effectively. Besides, by exploring the outcome of the predictive model, research findings are expected to go beyond the exploratory purposes by suggesting alternative measurement to calculate song rankings by leveraging social media sites. The conclusion of this thesis should not only shed new light on the music industry and social media relationship but also apply

to other industries in which people are also likely to publicize their thoughts on social media sites, creating a bombardment of valuable data for research purposes.

1.6. Chapter Outline

The remainder of this research will be presented as follow. The next chapter consists of a background of the music industry which encompasses the development of music industry over the years, the relationship between social media and music industry, lastly the concept of song rankings to explain the reasoning behind choosing this research topic. Chapter 3 explains the literature review and some main concepts explored in this research, which are text mining, machine learning, and social media analytics. Chapter 4 presents a research methodology in which a detail description of the whole process of text mining and analysis are thoroughly explained with the ultimate goal to answer main research questions. Afterward, chapter 4 describes the findings of this thesis before discussing the final results. Last chapter – chapter 5 – presents the conclusion and points out the limitation and research gap in order to make recommendations for future researches.

Chapter II: MUSIC INDUSTRY

The objective of this chapter is to provide background information of the music industry in order to give readers an overview guiding the interpretation of the following chapters. Thus, this chapter starts with the development of music over the years. Then, the relationship between social media and music are going to be presented in order to explain why the association between Twitter and song performance was chosen as a research topic, coupled with findings of previous studies of the same topic.

2.1. The development of music over the years

According to (Marketing Charts, 2017), American people spend around 4.5 hours per day listening to music. Music has been playing an inseparable role in our lives as people expend a great amount of time enjoying music on a daily basis. At the same time, the ubiquity of the internet and modern technology has changed substantially the way listeners consume music. Traditionally, back to 1920s, radio stations began broadcasting music that enables music lovers to enjoy music at their homes. Later on, during the 1970s, the first cassette-playing Walkman by Sony allows people to play cassette tapes of their favorite songs anywhere. Then 1990s marked a breakthrough of technology when internet invention significantly left an impact on every aspect of lives where the customer went from purchasing CDs in-store to paying for individual MP3. Fast forward to 2017, radio has been made almost obsolete, and digital music services like Apple, Spotify began to come around with downloading of music starting to give away (Throckmorton, n.d.). According to Nielsen's US Music 360 2017 Report, among 90% of music listeners, 41 of their time is spent listening to music streams.

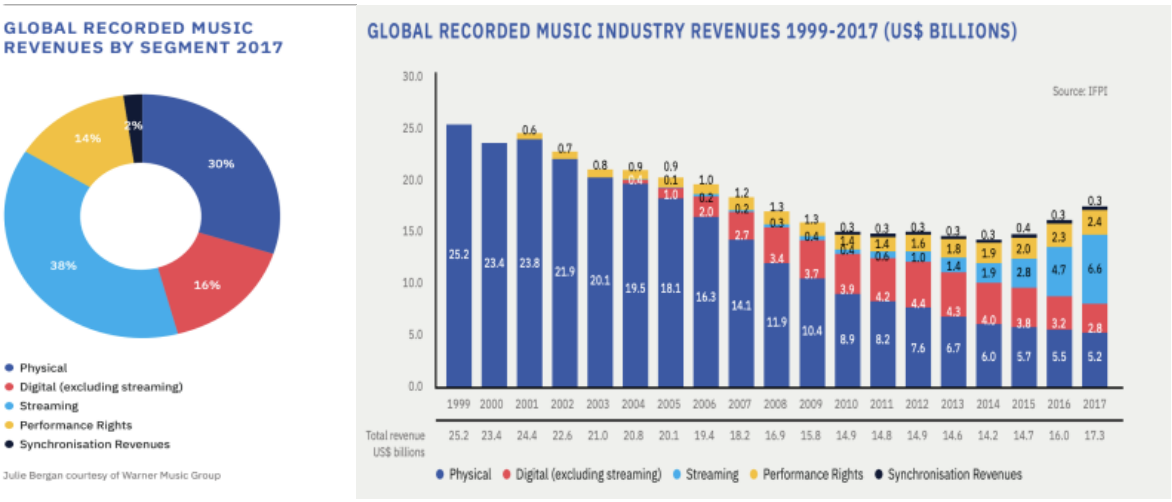


Figure 2.1 Global recorded music revenue by segment (IFPI, 2018)

As can be seen from the chart above (IFPI, 2018), the music industry has experienced an on-going decline since 2001. Up until 2014, the total revenue stream marked its lowest end in the history of the industry – 14.2 billion USD. However, as reported by IFPI, there was a slight increase in 2015 – around 3.4%. Later on, in 2016, streaming surpasses digital downloads; consequently, the revenues saw significant growth over more than a decade for the first time of 8.9%. By the end of 2017, 38% of total revenue was already occupied by music streams. Simultaneously, it can be clearly seen than social media usage has experienced massive growth. Not surprisingly, this growth brings in the significant integration of social networks and music platforms, offering a new way for music fan engagement. The relationship between social media and the music industry is discussed further in the next part.

2.2. Social media and music industry relationship

The booming of social media has created its way to emerge deeply into the music industry. It is widely believed that social media is highly integrated into the music industry, particularly after the introduction of streaming music in the 2000s. Nowadays, the internet has made accessibility to music in just a matter of seconds (Molla and Kafka, 2018). Social media is considered a platform where music lovers find out about new songs, new artists, and communicate with like-minded people. Music fans are given chances to directly correspond with their favorite artists by sending a private message or write on their public wall. On musician side, they have opportunities to reach out directly to their fans by engaging into a closer community with them and promote their images via various features of social media. However, along with the music digitalization, one of the huge drawbacks is that the industry is

suffering from decreasing revenue due to illegal downloads, peer-to-peer sharing – the cheaper gateway to access music. Therefore, numerous studies have been carried out to investigate the correlation between social media and song revenues. The exponential growth of social media and music streams has turned closed-format music outdated (Collins, 2013). Recent years, most streaming apps perform such a poor task in executing social functionality themselves; there is room for social networking sites like Tiktok, Youtube, Spotify to deliver a portfolio of social music experience. Figure 2 below visualizes the position of most common music streaming services and social networks according to their music-focused and social-focus levels.

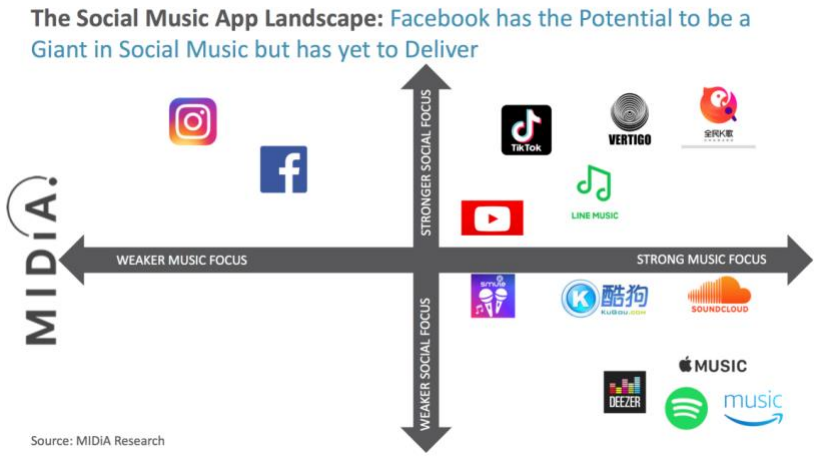


Figure 2.2 Social Music App Landscape (Mulligan, 2019)

Social media has changed so much over time and brought tremendous improvements in some areas in the music industry. According to Billboard (2018), 9 out of 10 social media users perform a music-related activity on their social media application. Besides, 41% like share and comment on artist postings and 29% share songs, albums, and playlists from streaming services. Among top social media, Twitter has experienced a considerable contribution of music to social engagement on the site with 51% users using Twitter to follow or get updates from music artists and bands (Crupnick, 2018).

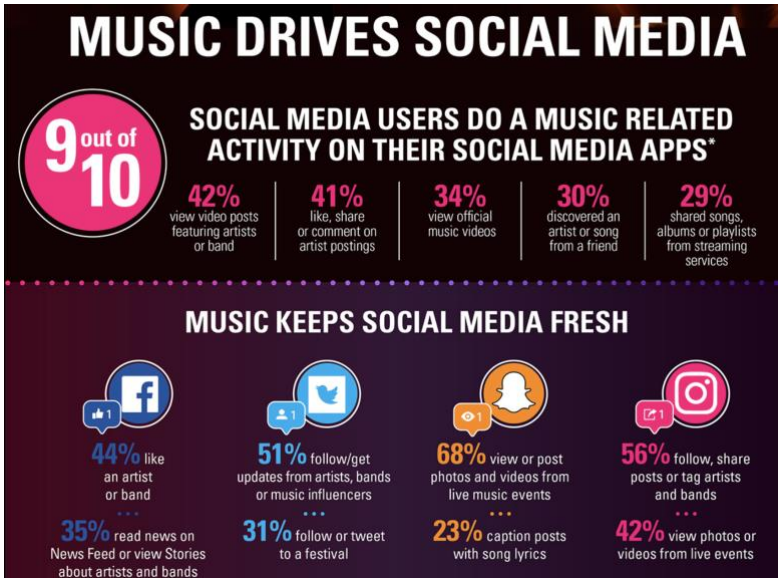


Figure 2.3 Music drives social media (Crupnick, 2018)

Twitter created a very different type of environment which created the most open communication forum between fans and artists and between fellow artists, giving artists, actors and the like a chance to let their true feelings and deepest thoughts out to their huge followers.

2.3. Twitter and song rankings

Record chart also called a music chart, is a ranking of recorded music based on specific criteria in a given time period (En.wikipedia.org, n.d.). Charts can be calculated based on any genre or geographical location, which are updated automatically every week. Music charts are meant to visualize qualitative song performance to the quantitative measurement for further analysis. There was no clear definition of a music chart back to the past. It is widely believed that the song's revenue is meant to be the definition of song popularity. Up until July 1940, the first music popularity chart was published by Billboard magazine. A variety of charts were released afterward, but the most popular chart, which is Hot 100, was not released until 1958. Hot 100 chart represented top 100 hit songs which were rated by single sales, radio airplay, digital downloads, streaming activity (including data from Youtube and other video sites) (En.wikipedia.org, n.d.). For decades, there were just physical retail sales and radio airplay to define song popularity, whereas nowadays, music listeners have plenty of ways to consume music from different sources, and each was created differently. Music is now being consumed on streaming services in more diverse ways, migrating from pure on-demand experience to a more diverse

selection of listening preferences such as playlists, radio…2016 and 2017 experienced high participation of streaming in music market which alters the way consumers engagement in music. Billboard thus frequently reevaluates song rankings measurement on how to incorporate streaming data into song charts. Streaming is given the highest weight in the chart, followed by radio play and digital sales in descending significance order, making up the three metrics of Hot 100's methodology.

Today it is not uncommon to use social media to share and discover new music. (Franklin, 2012) Mentions how "as music fans, most of us turn to Twitter or Facebook to keep updated about our favorite bands, whilst new tracks or videos will 99% get their launch on social channels. Furthermore, social media is where music audiences naturally congregate, forming their communities, and sharing their listening experience". These opinions and sharing are believed to be the most likely factors to demonstrate the performance of song among the online community. Regarding the strong relationship between social media and music, many researches were aimed to investigate further this relationship. Twitter as the main data source was exploited from different perspectives such as predicting the success of artist's upcoming release (Garzon, Luo, and Vazquez, n.d.), forecasting music sales revenue (Vossen, n.d.), understanding music listening behavior (49). Another popular aspect of the music industry, which is music charts, has not been investigated thoroughly yet. To the best of author's knowledge, only two previous researches by

However, another popular aspect of music, which is song performance was the research topics of only several papers. Among those, two researches by Zangerle (2016) and Kim & Suh & Lee (2014) aimed at investigating the relationship between Twitter and song performance. Kim (2014) initiated the trend of exploiting Twitter data in forecasting song performance, and Zangerle (2016) extended on this research by covering a longer time frame of 2 years, utilizing the time-lag feature to perform the analysis.

Music lovers want to let the world know what song they are currently listening by #nowplaying tweets, and even some music streaming platform like Spotify auto-generate hashtag #nowplaying once Twitter users share a song. Research by Kim (2014) is revolving around mining Twitter data by music-related hashtags #nowplaying to collect users music listening behavior over the course of 10 weeks in order to forecast Billboard rankings and hit music. The author used three different features to form the dataset, which are the song's popularity on Twitter, artist's

popularity on Twitter, and a number of weeks a song was in the chart. The results show that the number of daily tweets about a specific song obtained the highest correlation with song ranking. A support vector regression model which was built using those three features produced the best predictor performance with $r^2 = 0.75$. Furthermore, the research shows that top well-known songs (rank 1- 10) are likely to be forecasted with a precision value of 0.92 and a recall value of 0.88 (Kim et al., 2014). Another finding is that users music listening behavior on Twitter is highly correlated with general music trends and can be considered a channel to understand consumers music consumption patterns (Kim et al., 2014).

In the research of Zangerle (2016), #nowplaying was also chosen as access to tweets about music. What differs this study with the study of Kim (2014) is that it was implemented over the course of 2 years and it was solely based on Twitter data only without considering how long a song was in the chart. The result suggested that " Twitter and Billboard time series for tracks acquired a moderate correlation, which is influenced by a timely shift between those two." The paper also mentioned that solely relying on Twitter data for charts prediction tend to be highly error-prone while multivariate model incorporating Billboard and Twitter data is found to perform prediction accurately.

This dissertation also aims at discovering the power of Twitter in predicting song performance. What differs this study with previous ones is the inclusion of sentiment analysis, which none of the previous researches has ever done before in the music industry. Therefore using sentiment analysis of Twitter as a mechanism to explore the relationship between online chatter and song performance is still an open question. Besides, instead of restricting to only one hashtag #nowplaying, this study will exploit all tweets related to both tweets and artists by using hashtag (#) and mention feature (@). Additionally, under the time restriction, the limited number of data collected only allows the song ranking prediction as a subgroup of Top 10, Top 20…instead of exact position like prior researches (this will be discussed in chapter 4 methodology). Next chapter – Literature review explains all concepts used in this dissertation.

Chapter III: LITERATURE REVIEW

This chapter is aimed to give readers an overview of concepts related to the research undertaken. It starts with a brief introduction of text mining, followed by necessary steps to perform text mining. In addition, machine learning techniques are presented with the intention to choose the most suitable algorithms for the undertaken research. Lastly, Twitter analytics part explains the reasons behind choosing Twitter as the primary data source. In addition, two primary Twitter variables are discussed further under the context of some papers with the same data source as Twitter.

3.1. Text mining

Data mining has experienced rapid growth in recent years because of immense advances in hardware and software technology, which led to the emergence of different type of data. This is extremely true in the case of text data where the widespread popularity of social network platforms has allowed the quick creation of huge repositories of various kinds of data (Aggarwal and Zhai, 2012). 80% of available data is unstructured, while only 20% is structured data (Salloum, 2017). Thus the ubiquity of text mining provides a great data source for any research purpose. Text is basically everywhere in our daily lives. Underlying the search engines or daily web browsing that people use every day is an enormous volume of text-oriented data science. This user-generated content is specifically informative in a business setting where customer feedback usually takes the form of text. Text mining is therefore required to extract insights out of these unstructured texts.

Text mining or knowledge discovery from the text (KDT) was first introduced by Fledman et. al. It refers to the process of extracting interesting and significant patterns from natural language source instead of structured data (Gupta, 2009). It requires a demanding skill of processing data to transform unstructured data from different data sources to an easy and structured format for further analyze. Human has the ability to distinguish and apply linguistics patterns to text, and more importantly, human being can easily tackle slangs, idioms, and contextual meaning (Gupta, 2009). However, when it comes to process data at a large volume and high speed, human being with good comprehension is merely insufficient to provide the result without compromise the quality. Regardless of the difficulty in collecting data,

text mining is treated as an intelligence system which extracts insights from amorphous sentences and improper words then transform into valuable insights to make particular suggestions.

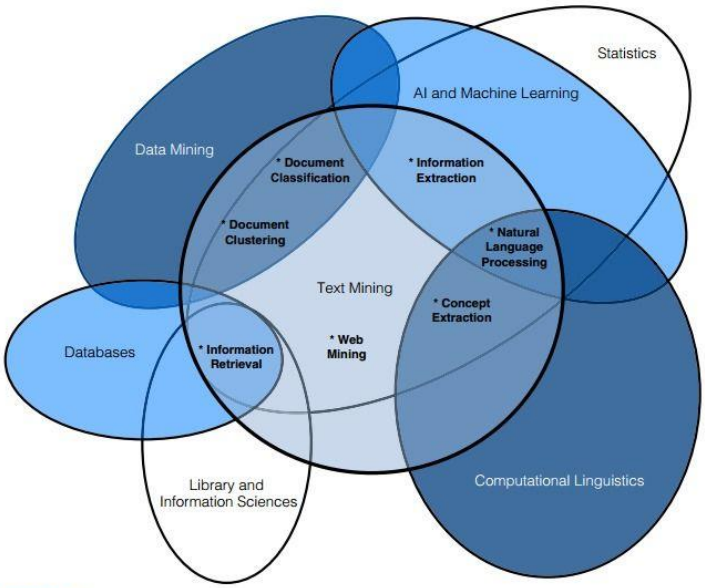


FIGURE 2.1
A Venn diagram of the intersection of text mining and six related fields (shown as ovals), such as data mining, statistics, and computational linguistics. The seven text mining practice areas exist at the major intersections of text mining with its six related fields.

Figure 3.1 A Ven diagram of text mining and six related fields (Talib, 2016)

The purpose of text mining as discovery and extraction insights non-trivial knowledge from unstructured text (Kao and Proteet, 2010) encompasses everything from information retrieval, text clustering and document to natural language processing and concept extraction. Figure 4 is a Ven diagram describes the intersection of text mining and six related fields which can be exploited by text mining techniques.

Text mining has been also utilized in code mining of other fields such as software engineer. Those techniques follows a data-driven approach to recover names by searching in a large corpus of data (Tran H., 2019) and to understand the code semantic (Tran N., 2019, and Nguyen H., 2019).

The primary goal of this undertaken research is to transform Twitter data into minable data for analysis by using Natural Language Processing (NLP) and analytical methods. Natural Language Processing is a sub-field of computational linguistics and text mining which aims at transforming the free text of natural language into machine-readable format and display in a statistical approach (Kao and Proteet,

2010). "It is one of the hot topics that concern about the interrelation between the huge amount of unstructured text in social media" (2). According to Aggarwal and Zhai (2012), streaming data is more challenging to mine than others because they need to be processed in the context of one-pass constraint. In other words, it might be difficult to store data offline for analyzing, and it is necessary to perform the mining task continuously as the data comes in. The task is more challenging when it comes to mining text in social media where the text contains poor and non-standard vocabulary, which could subsequently hinder the performance of sentiment analysis. However, Natural Language Toolkit (NLTK) written in Python programming has been developed and expanded with the primary goal to tackle unregular text like Twitter data. It consists of most common algorithms such as tokenizing, part of speech tagging, stemming, sentiment analysis, topic segmentation, and named entity recognition (Kao and Proteet, 2010). NLTK is aimed to help computer analyze and understand writer text in various format. In this dissertation, it is utilized to perform sentiment analysis of tweets on Twitter.

Similar to data mining, text can be mined in a more systematic and comprehensive way, but the generic process performs five following basic steps:

- *Data Gathering*: In this step, unstructured data are collected from different sources, either external or internal. Data gathered can be presented in various file formats such as plain text, mixed text, numerical, web pages, etc.
- *Data Preparation*: Unregular data would never come in handy without the data preparation step. Hereby different Natural Language Process techniques such as tokenization or part-of-speech tagging were used to clean data and transform it into an easy-to-read format in order to build an input variable set.
- *Data Analysis*: The most crucial part of the whole mining process is to extract insights from data. The results of previous steps make clear some underlying patterns from data. These patterns will, in turn, be analyzed by data analytics tool, resulting in information which can be used for effective and timely decision making and trend analysis.

3.2. Machine learning

Complexity and volume of data make it hard to for human to process and comprehend all meanings behind it (Zangerle, 2016). Machine learning can be used to find interesting and meaningful information in an automated way. Arthur Samuel (1959)

asserted that machine learning is a field of study that gives computers the ability to learn without explicitly programmed. Machine learning, according to Tran (2019), machine learning a paradigm that performs algorithms from past experience to improve future performance. The whole process revolves around the observations of data, patterns recognition, and better-making decisions in the future based on given examples.

There are two types of machine learning which are supervised and unsupervised learning algorithms. *Unsupervised learning methods* are techniques for finding hidden structures under unlabeled data. In-text mining concept, clustering, and topic-modeling are the most used algorithms. Cluster is aimed to segmenting a collection of documents into a different cluster where documents in each partition are more similar to each other than those in other partitions. In topic modeling, on the other hand, each topic can be represented as a probability distribution over words, and each document is expressed as a probability distribution over topics (Allahyari, 2017). *Supervised learning methods*: is a machine learning technique to infer a function or learn a classifier from training data in order to implement predictions on new data. Each example has a label which tells the system what the correct output value is. The labels then tell which category each example belongs to.
The main task is to define patterns then correctly map input match the output. By these algorithms, researches explore data with the hope to draw implications from hidden structures of unlabeled data set.
Supervised learning methods are classified into two types: Regression and Classification. Regression techniques are used to predict continuous value, numerical responses based on previously observed data. On the other hand, classification is aimed to predict discrete responses. It is often used when data are categorized, grouped, and separated into a specific partition. Classification models, therefore, predict the categories that input data belongs to, which is called predictive modeling. Prediction more generally means estimating an unknown value (Provost and Fawcett, 2013). This value could either be value in the future, present, or in the past. The framework is shown in the figure below: (a) During training, a classification model is generated by machine learning algorithm with a set of features that are extracted from labeled input data via feature extractor. (b) During prediction, unlabeled data is fed into the classification model via the same feature extractor. The trained classifier model has the eventual goal of predicting labels of these input data based on these features. (Bird et al., 2009)

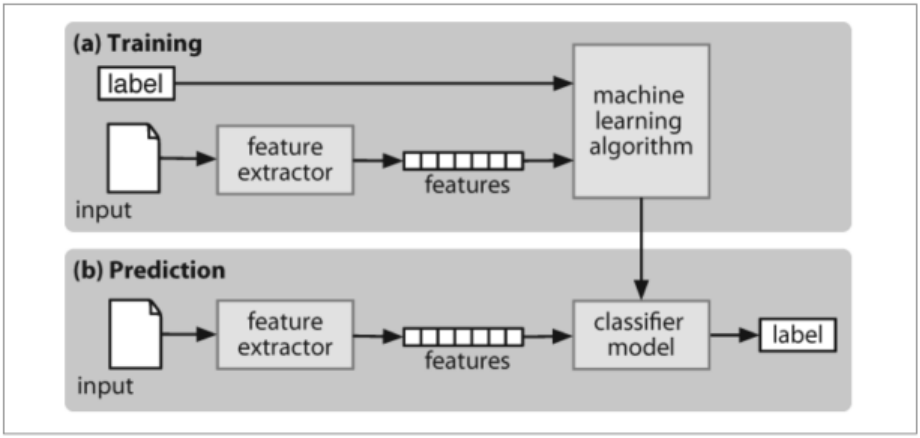


Figure 3.2 Supervised classification workflow (Bird et al., 2009)

In this dissertation, the song ranking as the target variable is a discrete variable and eventual goal of the research is to forecast them based on independent variables set of Twitter data. Thus, the author found supervised classification technique is the most suitable approach.

3.3. Twitter Analytics

It is undoubtedly that the explosion of social media has made enormously radical changes to our lives. According to Fan (2014), 91% of online adults use social media regularly, and a regular user spends more than 20% of their time on social media sites. A study by (Salloum, 2017) asserted that social media platforms are providing a great space for individuals to facilitate interactions and share their views and opinions. People, therefore, are more connected to each other despite the different nationalities, different customs because those views allow them to understand a particular person's activity, their emotions, and areas of personal interest to some extent. As the number of users on social media sites continues to increase, so does the amount of user-generated content. Social media analytics thus have transformed these data into valuable insights for academic researches. Social media analytics, according to Fan (2014), "is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize and visualize social media to facilitate conversations and interactions to extract useful patterns and intelligence." Social media analytics is a "process of gathering data from social media platforms and analyzing data using diverse analytical techniques to extract vital insights, which can be used to make data-driven business decisions" (Bali, Sarkar and Sharma, 2017)

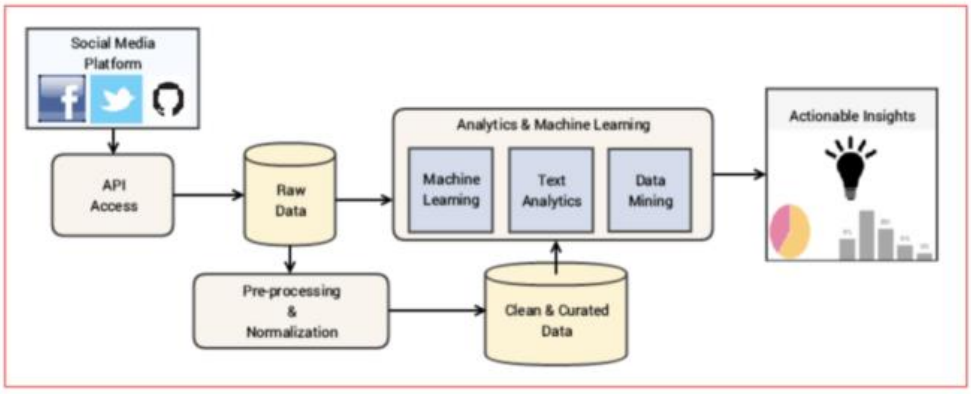


Figure 3.3: Social media analytics workflow (Bali, Sarkar and Sharma, 2017)

Figure 6 demonstrates a good grasp of the essential steps of a typical social media workflow. The main stages are basically similar to the text mining process. However, data accessibility requires either APIs which is provided by social media platform or unofficial mechanisms such as web crawling or scraping. Crawling data is indexing information on the web pages using bots (crawler) while scraping data is merely retrieving information from any source (Jha, 2012). Besides, text collected from social media is usually comprised of many components that make the pre-processing and cleaning steps even more difficult.

Among common social networking platforms, Twitter is considered most used sites for research purposes in comparison with others such as Facebook, Instagram. Zimmer, Proferes, and Nicholas (2014) found that 38% of Twitter research was carried out in the computer science field with information science (21%) and communications (14%). This preference could be explained by several following points. Firstly, Twitter is popular for its simplicity and accessibility, which provides a large amount of tweet on a vast number of topics. Besides, Twitter has an embedded social network which determines how content disseminates among users. In particular, Twitter provides an application program interface (API) structure that makes data collecting process much easier (Huaxia, 2015). API (Application Programming Interface) is a set of subroutine definitions, communication protocols, and tools for building software (En.wikipedia.org, n.d.). Twitter APIs allow companies, developers, users programmatic accessibility to Twitter data (Help.twitter.com, n.d.). There are two types of standard (free) Twitter APIs, which are REST APIs and Streaming APIs. REST APIs is a web service that follows a request-response pattern which is suitable for users who want a snapshot of data that does not change frequently. Streaming APIs meanwhile maintain a constant connection

that continuously sends updated data to the user until connection is terminated (Snowflake Software, n.d.). Besides, Twitter makes it easier to collect data from topic-based groups; moreover, metadata such as geographical data like precise longitude and latitude from which tweets are created is also included in the scraped data that delivers a wide range of users information useful for research purposes.

Given the benefit of being able to analyze Twitter data is to understand the sentiments of what users are currently discussing on the Internet, a review of literature has shown that diverse topics such as stock market movements (Bollen, Mao, and Zeng, 2011), deep learning for depression detection (Orabi, n.d.), the USA presidential election results which are all done by Twitter analytics. Though unstructured Twitter data requires an entire effort to be transferred into an easy-to-read format, there are two main predictor variables extracted, which are volume-related variables and sentiment-related variables. In order to have a better picture of the importance of these variables in academic research, several researches in different contexts, different industries are presented as follow.

- *Volume variables*

It is the most common type of social media variable, which at the same time is found as a good predictor variable. The movie industry has been exploited from various perspectives in which Twitter is the main data source. One of the studies initiated the research trend of Twitter data is the research by Asur and Huberman (2010) in the movie industry. In this research, the combination of volume (the rate at which movie tweets are generated) and sentiment content in the tweet as valence variable can be an effective indicator of sales revenue, and the results are considerably better than those produced by information market (Asur and Huberman, 2010). Rui (2011) extended on this research by filtering tweets then classifying them into four mutually exclusive categories: intention to watch movies, positive, negative, and neutral. The result shows that intention tweets ratio, the total number of tweets, tweets from users who have a high number of followers and ratio of positive tweets all have a significant and positive influence on movie-box office revenues.

In the context of the healthcare industry, the predictive power of Twitter is not as strong as in music or movie industry but is still useful to some extent as a monitor. Achrekar (2011) presents Social Network Enabled Flu Trends framework "which monitors messages posted on Twitter with a mention of flu indicators to track and predict the emergence and spread of an influenza epidemic in a population." Similar

to a majority of researches, the volume of flu-related tweets is highly correlated with the number of influenza-like illness cases reported. A correlation was also found between suicide-related tweets and actual suicide rates in research by O'Dea (2015).

- *Sentiment-related variables*

In the research of Rui (2011), he found that sentiment of tweets mentioning a movie for each week, an explanatory variable to capture the valence effect of Twitter on movie box office revenues is the main driver to predict total revenue not volume variables like previous studies. However, in the study of Asur and Huberman (2010), while volume-related variables produce a high accuracy in forecasting box office revenues, the inclusion of sentiment was found to have a small effect to the total revenues and accuracy of prediction only slightly increased. In another research using Twitter sentiment to predict the winner of the US 2016 Presidential Election of Tunggawan and Soelistio (2016), presidential nominees are predicted by looking for candidates with the most predicted positive sentiment. As shown above, sentiment-related variables were found in different topics; however, none of the articles of the predictive power of Twitter data in the music industry has performed sentiment analysis. However, it certainly seems relevant to include a public sentiment on social media in predicting song performance. As observed from the real-life context, high-performance songs are indeed favored by music lovers; therefore, are likely to be positively mentioned on social media than lower-ranked songs.

To sum up, this dissertation will exploit NLP to perform sentiment analysis as the primary text mining techniques. In order to meet the targeted goal of predicting song ranking based on Twitter data, supervised classification was found as the most appropriate machine learning algorithm. In the subsequent chapter, Methodology, the decisions taken during research design and all three main steps of text mining which are data collection, data preparing, and most importantly data analysis part will be presented in detail.

Chapter IV: METHODOLOGY

The objective of this section is to guide the readers through methodological decisions undertaken in this process. Section 4.1. explains undertaken research design, which includes research purpose, approach to theory development, and research strategy appropriate to guide the whole research to answer the main research question. Section 4.2 describes the data analysis workflow in which the CRISP-DM model was employed to guide readers through the whole analyzing process.

4.1. Research Design

4.1.1. Research Purpose and Approach to Theory Development

According to Saunders, Lewis, and Thornhill (2009), the way research questions are formulated results in explanatory, exploratory, or descriptive or evaluative answers forms the main purpose of the research.

Exploratory studies are often employed when detail is presented limited to the observed phenomenon of interest. This type of research is often aimed to find out "what is happening to seek new insights, to ask questions and assess phenomena in a new light" (Saunders, Lewis and Thornhill, 2009) therefore it is a good fit for those who want to explore or make clear the problem without knowing surely the research outcomes.

Descriptive studies are a piece of exploratory or explanatory research. This type of study is suitable for researches answering research questions such as why the phenomenon comes to being or why should it be considered to be necessary. This study often needs inferred conclusions upon the findings.

Explanatory research, on the other hand, is meant to establish causality between variables. The purpose of the undertaken research is to predict song rankings based on Twitter data, but it is difficult to prove a causal relationship. Since the high ranked song is likely to be tweeted, which consequently makes it much more popular and favored, leading to higher ranking in the chart. Therefore, this dissertation is thus designated as an exploratory study. This is not used to give new conclusive evidence

but rather shed new light on problems and help the understanding more efficiently in a new light.

In addition to this, exploratory research usually associates an inductive approach to theory development. Inductive studies involve the research for a pattern of observations and gain awareness of a topic of interest in order to build a theory (Saunders, Lewis and Thornhill, 2009). At the same time, inductive studies also allow researchers to formulate research questions based on existing theory. Provided that social media provides a great data source for researches about people's behaviors, it is going to be explored upon its power to predict song performance in this study.

4.1.2. Research Strategy

According to Saunders, Lewis, and Thornhill (2009), Research strategy enables researchers to answer particular research questions and sub-questions which is guided by research questions, the extent of existing knowledge as well as the amount of time available. There are eight different strategies that will enable the author to answer research questions and meet the goal of research: experiment, survey, case study, action research, grounded theory, ethnography, archival research.

The undertaken research follows grounded theory strategy because the research starts with collecting data from Twitter and Billboard, then analyzing observations obtained in order to build a theory upon the found results. Grounded theory (Saunders, Lewis and Thornhill, 2009) is particularly helpful for research which aims at explaining and predicting behaviors, "the emphasis being upon developing and building theory." It is often used in exploratory studies in which data collection starts without the formulated theory framework. The observation obtained from data analysis will lead to the generation of predictions, which are then tested in further explanations.

Regarding the time horizons of the research topic, a longitudinal study is expected to produce the best result since the longer time frame allows researchers to exercise a measure of control over variables being studied, provided that the research process itself does not affect the result. The previous study by Zangerle (2016) asserted that the study of Kim (2014) was limited due to the short analysis period of ten weeks thus their study is based upon 2-year-collected data, covering time frame substantially longer previous researches and utilizing time series to perform analysis.

This study follows a cross-sectional time horizon since data collection is limited over a short period of 2 weeks. The implication is that song ranking charts will be predicted as a subgroup instead of exact rank position (which is elaborated more on testing and modeling part).

4.2. Text mining analysis

4.2.1. Data Collection

The undertaken research requires data from two different sources gathered over the same time frame of two weeks from 30th June to 13th July. Primary data was directly crawled from Twitter to build an independent variable set. Besides, the collection of songs rankings will be extracted manually via Billboard Hot 100 chart, which is treated as a dependent variable. In addition, another variable which is the number of weeks the song was in the chart was also collected manually via Billboard.com. The inclusion of this variable increased prediction accuracy as being found in the research of Zangerle (2016) and Kim (2014). Therefore, it will be employed in this dissertation to find if any positive results could also be found.

Twitter data

The chosen software to perform Twitter data collection is Python. According to Gupta (2017), Python is well-known for its easy code readability and solid lines of codes. It is also “a collection and code stack of various open-source repositories," which can be exploited for different purposes like NLTK or machine learning (Patel, 2018). This research collects Twitter data via Streaming API, which is easy access to live real-time tweets at the point of data collection, why offering low latency (Snowflake Software, n.d.). Tweets which were related to songs and artists in the Billboard Hot 100 song were updated continuously within two weeks. Many song titles which consist of general phrase result in the collection of irrelevant tweets. For example, by using hashtag #talk, #sucker, or #wow for each song Talk, Sucker, Wow respectively generated numerous unrelated tweets that did not discuss anything about music. Therefore, familiar-titled songs were omitted from the list, so data quality will not be affected by noisy data. This was easily done by manually check each song title in Hot 100 song list. Out of the original 100 songs, only 37 songs were finally chosen.

Matching feature artists name, on the other hand, is more complicated since there are many formats in use. For example, Ed Sheeran ft Justin Bieber can be listed as either "Ed Sheeran ft Justin Bieber," "Ed featuring Justin," ”Ed and Justin” which is too complicated to finalize the hashtags with the artist name. Therefore, to simplify this problem, the author instead used mention @ symbol to extracting tweets that mentioned the main Twitter account of the artist such as @edsheeran or @justinbieber. This paper will eventually make sure data used clean and clear, disambiguated by choosing appropriate keywords for both artists and songs, finalizing a list of hashtags and mentions ready for next steps.

The figure below illustrates the main steps in Twitter data collection and sentiment analysis process, which were all done by Python. Firstly, Twitter streaming API provides access for the crawler to collect live updated tweets. Crawlers permit fine-grained control over where to look, which links to follow, and how to organize the results (Bird et al., 2009). Consequently, the outcome of data collection regarding retrieving real-time tweets ended with 1.017.627 tweets. Next, data collected were saved in Data storage as a JSON format, which consists of ID-value pairs and array data types. Then, the whole practice of data processing was to transform raw data into an easy-to-understand format, simultaneously data was also sent to the NLTK library for sentiment analysis. Finally, data as a csv format was ready to be transferred into Rapid miner for data analyzing step.

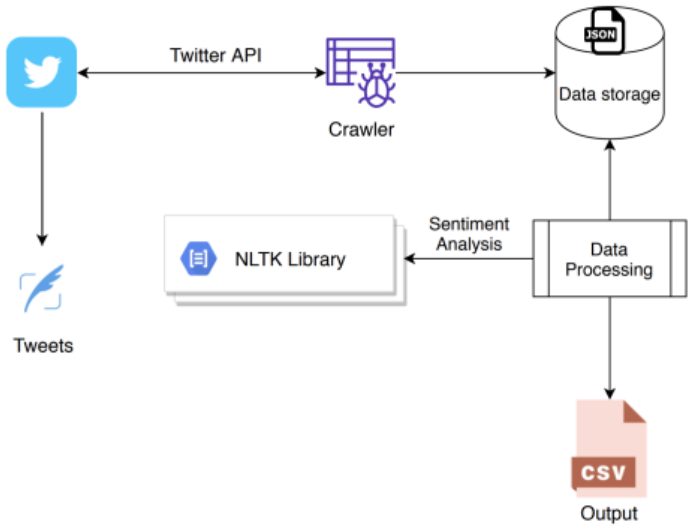


Figure 4.1 The architecture design of workflow

Billboard data

Billboard ranking chart can be easily collected via Billboard website. Pursuing longitudinal study, previous researches by Kim (2014) and Zangerle (2016) collected a large amount of data which allowed the forecast of exact song position in the chart. Meanwhile, this dissertation is carried out in an academic context with time constraints, resulting in a limited amount of data collected. This lack of data might hinder the performance of supervised classification algorithms. Therefore songs are classified into subgroups of ranking instead of their precise chart position such as Top 10, Top 20, etc. (This will be discussed further in 5.1 Results part).
Note: Songs which no longer stayed in Hot 100 chart in week two were omitted from the dataset.

4.2.2. Data Analysis

Once data was collected and transformed into one CSV file, it is ready to be analyzed. Rapid Miner was chosen to perform data analysis. Rapid Miner provides 99% of advanced analytical solution through template-based frameworks that do not compromise delivery and quality by nearly eliminating the need to write code (En.wikipedia.org, n.d,). Rapid Miner Studio Free Edition represents a reliable and easy-to-use tool which allows entry-level data scientists or even researchers with no background in data science to get the most out of any given dataset easily. Analyzing data is going to be described based on an updated model of Cross-industry Standard Process for Data Mining (CRISP-DMDM) to guide the whole process systematically. The process model consists of 6 phases, which are well structured and defined: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment.

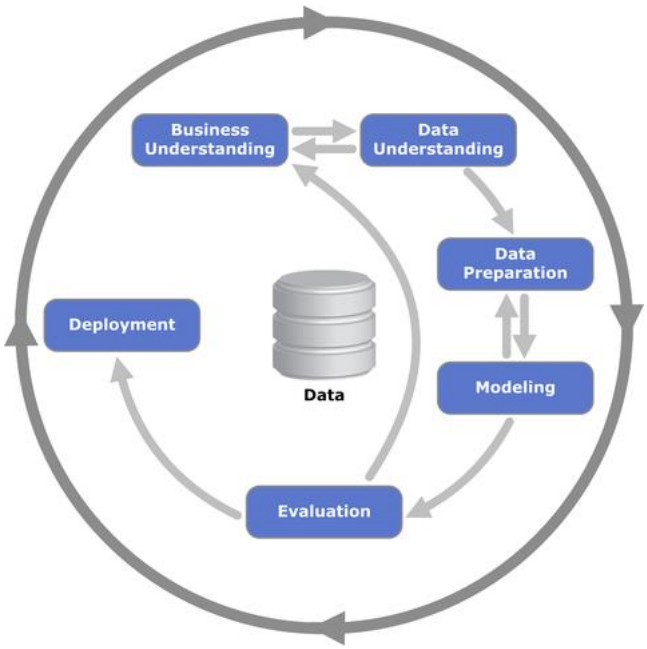


Figure 4.2: The CRISP-DMDM Data Mining Process (Provost and Fawcett, 2013)

4.2.2.1. Business understanding

The first phase defines the main objectives of the investigation and represents some familiar aspects of the whole process (Wirth and Jochen, 2000). Chapter 2 covers the broad introduction of the music industry, in which the connection between social media and the industry was highlighted. The fast growth of streaming music has been contributing to the integration of social media in music industry. Hence, leveraging the impact of online chatter on the industry is expected to add values in better managing social networking sites such as Twitter, Facebook, initiating appropriate online marketing strategy to promote songs and improving social interaction.

4.2.2.2. Data understanding

The second stage starts with an initial data collection then get familiar with data in order to discover the first insights into the dataset (Wirth and Jochen, 2000). As discussed in the Data Collection part, tweets which contain hashtags about songs and mentions about artists are retrieved for research purpose. Hashtag (#) plus the title of song such as #youneedtocalmdown allows the crawler to collect all tweets at that point of time which contain #youneedtocalmdown. Besides, symbol sign (@) is often used to refer to the user account on Twitter. Twitter API creates a condition for researches to collect users' information from the tweet itself, such as time post,

geographical location, number of retweets per tweet. In this dissertation, only the content of tweets was retrieved for further analysis. Billboard data, on the other hand, is easily perceived due to its simple characteristics. Number of weeks a song was in the chart indicated for how long a song stays in Hot 100 music chart. Besides, song rankings as the target variable are grouped into different classes such as Top 10, Top 20, Top 30…

4.2.2.3. Data preparation

This phase covers all necessary activities to construct final dataset from initial raw data (Wirth and Jochen, 2000). The more noisy, irregular, and incomplete data, the more difficult it is to convert them into minable data. Even though the number of characters per each tweet is only limited to 140 words, tweets are composed of not only regular words but also emoticons, URLs, hashtags, mentions, or slang words. Thus, it is a demanding task when it comes to Twitter data preparation. Due to the large volume of tweets within two weeks, Python was used to process data then finally congregate into a well-formatted final dataset.

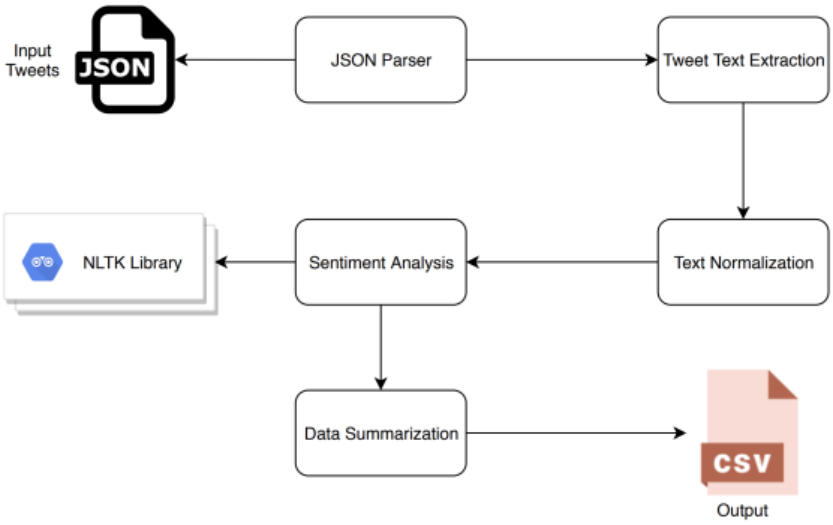


Figure 4.3: Subprocess of text mining workflow

Figure 8 illustrates the main steps in Data processing stage, which is a subprocess of Text mining workflow (Figure 7). Data preparation is to stop at Text normalization step while data processing encompasses data preparation and sentiment analysis and finally data summarization to transform data into CSV format. As can be seen from the workflow, the only text was extracted from original Json input tweets. By implementing some data cleaning steps (text normalization) such as removing stop words, emojis, normalizing all words into one case instead of leaving them all in

different formats, etc. The normalized text was then sent to the NLTK library for sentiment analysis, and once it is done, data summarization will aggregate all text and export into CSV file ready for thorough further analysis.

The final dataset contains the following variables.

- ***Number of tweets about songs***: This variable is calculated by the end of each week by calculating total tweets containing hashtag (#) and song title
- ***Total tweets about artists***: This is acquired by computing tweets covering mention @ and the name of artists
- ***Number of weeks a song was in Hot 100 Chart***: This variable is easily be collected per week via Billboard website.
- ***Sentiment-related variables*** are calculated on tweets related to songs only. Each tweet has three different scores: positive, negative, and neutral score, and these values fall into the interval 0 to 1 (Tran, 2016). Moreover, for every tweet, a sum of positive, negative, and neutral equals to 1.

    - *Average positive score*: $\frac{Total\ \ positive\ scores\ of\ all\ tweets}{Number\ of\ tweets}$

    - *Average negative score*: $\frac{Total\ negative\ scores\ of\ all\ tweets}{Number\ of\ tweets}$

    - *Number of positive tweets*:

    A tweet is considered positive if its positive score is higher than the other two. If positive and negative score is equal, tweets are defined as neutral tweets.

        - *Number of negative tweets*: Similar method is applied to calculate number of negative tweets.
- ***Song ranking***: is collected directly via Billboard website Hot 100 Chart. Songs are then grouped into different subgroups of rankings: Top 10 – Non-top 10, Top 10 – 20 – 100, Top 10 – 30 – 100, Top 10 – 50 – 100 in order to see which classification will provide the best prediction result.

Once the data combination came to an end, the dataset was once again checked for missing and invalid values, duplicate records. The outcome of this phase in one single file containing all data collected within two weeks from 30th June – 13th July in which song ranking is regarded as dependent variables and the rest as independent variables for prediction purpose.

4.2.2.4. Testing and Modelling

In this modeling step, the Pearson Correlation was applied to investigate the contribution of each independent variable in predicting the ranking chart. Next, the main purpose of this research is to build a predictive model of Twitter data to song ranking. The final dataset is divided into two parts: 2/3 for training data and 1/3 for testing. The ratio of splitting data was decided based on trial and error method since this ratio gave a better prediction result compared to other common ratios 80:20 and 75:25. Even though, previous studies noted that data mining algorithms perform better on larger training dataset, models with too many training data are prone to overfitting problem. Overfitting is the tendency of data mining techniques to generate models which tailor to training data without generalization to previously unseen data (Provost and Fawcett, 2013)

Different supervised machine learning algorithms are employed to preform prediction: Decision Tree, Neural Network and knn which were then evaluated and the one which yields highest accuracy level will be chosen to formulate the final model (Reiman., 2018)

- *Decision Tree*: is one of the most widely used and practical supervised machine learning method. It builds models in the tree-like graph, which breaks down data into subsets while at the same time, an associated decision is incrementally built (Tran., 2019). Each node represents an attribute, each brand defines a rule, and each leaf represents an outcome which altogether represents classification rules.
  There are several commonly used decision tree algorithms such as CHAID, CART, ID3, and C5.0 that differs in node structure, how to perform the splits, and when to stop splitting. Handling well both continuous and categorical data and missing values which is suitable to generate classification trees, C5.0 model is selected to build the model. C5.0 uses gain_ratio for its splitting process; thus, gain_ratio is set as the parameter of decision tree in this research.

- *Neural Network*: the main idea behind a neural network is inspired by the way biological nervous systems work. It is aimed to stimulate densely interconnected processing units inside a computer (Woodford, 2019). The

advantage of neural network is that it learns all by itself without programming it to learn explicitly. A typical neural network consists of input units, hidden units, and output units.

- *kNN*: k nearest neighbors at its most basic level classifies data based on similar traits they all assemble. Knn is widely used, particularly in classification problems due to its easy implementation and understanding, quick calculation without compromising the outcome predictive power. The "k" value denotes the number of nearest neighbors the model will consider. K value is usually odd to prevent tie situations. There is no formula of calculating the best value for k; it is instead picked manually by trial in order to get the best possible fit for the dataset. In this case, k value is set as 5.

- *Naïve Bayes:* is a set of supervised learning algorithms based on the assumption of conditional independence over training data set (Tran., 2019). Naïve Bayes works well with a large dataset in a complex context. Its advantage is to require " a small number of training data for estimating parameters necessary for classification" (Tran., 2019). Naïve Bayes even though produced an average prediction with an accuracy level of around 60% was not exploited in this study since it requires the independence between predictors.

4.2.2.5. Evaluation

The performance of predictive model is evaluated based on the accuracy rate and confusion matrix. Accuracy is a commonly used evaluation metric in data mining since it reduces classifier performance to a single number and is easy to measure (Provost and Fawcett, 2013). Best performance model has an accuracy level of 1.0, which equals to 0 error rate model whereas the worst is 0.0.

$$\text{Accuracy rate} = \frac{Number\ of\ correct\ assessments}{number\ of\ total\ assessments}$$

This study works with multiclass classification in which the whole dataset is divided into different small groups based on the song ranking variable. This often results in an imbalanced dataset where classes are not equally presented. With imbalanced dataset, it is pretty easy to acquire a high accuracy rate without actually making useful predictions (Nabi, 2018). Therefore, in this study, accuracy rate merely is not

sufficient as evaluation metrics. Accuracy rate, according to Provost and Fawcett (2013), is naive and had some well-known issues. To have a better evaluation of predictive performances, the confusion matrix is thus taken into account to summarize and visualize classifier performances.

The whole process of modeling, testing, and evaluating model is performed using Rapid Miner as illustrated as figure below.

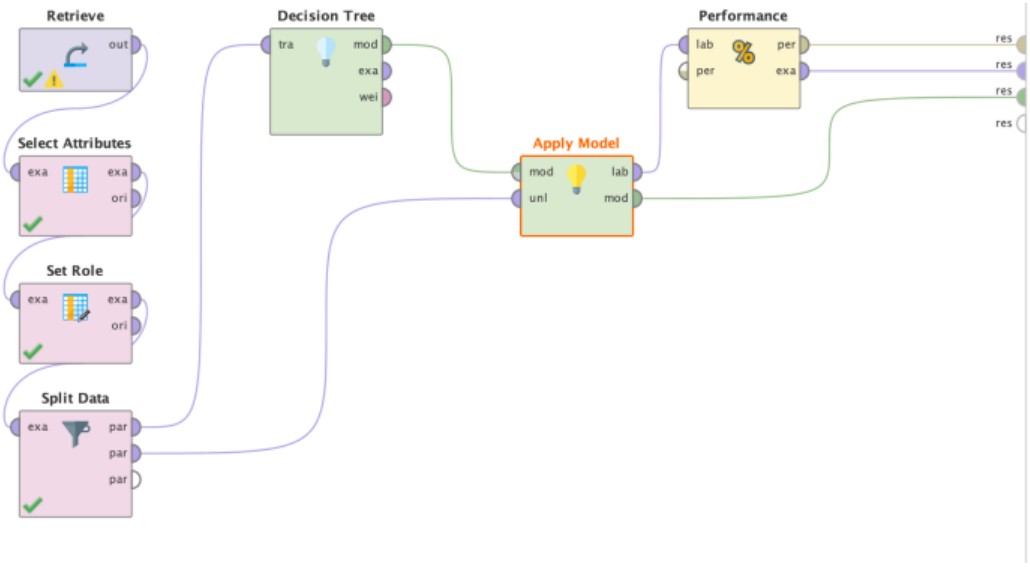


Figure 4.4: RapidMiner - Modelling, Testing and Evaluating process (Bird et al., 2009)

Chapter V: RESULT AND DISCUSSION

5.1. Results

This chapter is aimed to guide readers through the findings of this research with the ultimate goal to answer the main research questions. The results are designated into three parts. Firstly, in order to choose the most suitable target group classification, different subgroups of ranking are tested on the performance of predictive models. Besides, even though a combination of different variables normally performs better than a single variable (Chong et al., 2016), it does not necessarily evidence that all variables make a contribution to the models. Therefore, correlation matrix was applied for all independent variables with the intention to identify the relationship between each predictor and target variable. Lastly, the models built up by a combination of Billboard and Twitter independent variables were put in comparison with prediction made solely by Twitter data. Additionally, in order to contextualize, an attempt to relate findings of predictive model with the actual event occurred at the time of data collection is also going to be made. At the same time, all predictive models are compared in the level of accuracy and confusion matrix.

Choose the most suitable song ranking classification

Class 1: Top 10 – Non-top 10
Class 2: Top 10 – 20 – 100
Class 3: Top 10 – 30 – 100
Class 4: Top 10 – 50 – 100

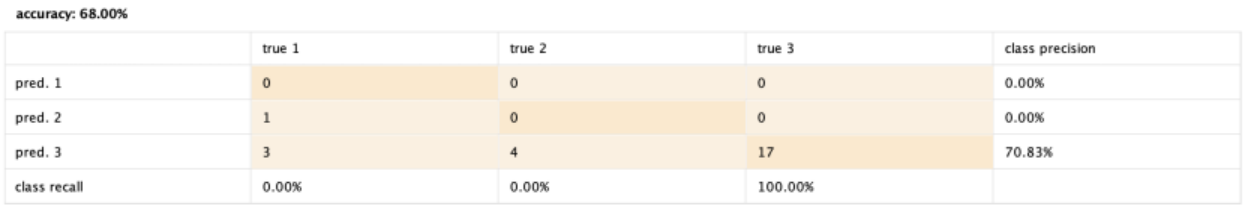
Time and date restrictions of undertaken research do not allow exact song ranking position; therefore, songs are ranked into subgroups instead. Different classes above are going to be tested under the performance of prediction; then the best performance class will be chosen based on accuracy rate and confusion matrix. All independent variables dataset is chosen to perform predictive models in this task. The following table illustrates the performance of prediction of each class.

Class	Decision Tree	Neural Network	kNN
Class 1	92	88	84
Class 2	64	68	52
Class 3	80.77	61.54	61.54
Class 4	56	56	64

Table 5.1: Accuracy rate of different supervised classification algorithms

It is noticeable that all predictive models in all different group classification generate moderate prediction accuracy level, which is above the baseline of 50% between 52% and 92%. Class 1 with simplified target group of top 10 and non-top 10 produces the highest accuracy rate, which is totally reasonable because these groups have quite clearly different characteristics, making it easier to predict which song belongs to which group. This occurs merely because of binary target option instead of better prediction algorithms for target variable. Among the rest, class 3 generally performed better than others because this classification probably allows supervised classification algorithms to easily generalize and classify songs into different groups based on their distinct characteristics.

Target group of Top 10 – 20 – 100, on the other hand, leads to the imbalanced dataset, thus resulting in overfitting problem. This can be explained by the fact that predictive models are likely to predict songs ranked 51st – 100th than top 10 and top 20 due to its superior number of songs. As can be seen from confusion matrix, while last group prediction produced relatively high-class precision of (70.83%), first and second group was mispredicted and was even not predicted. Probably, predictive models find it difficult to generalize patterns of these first two groups to categorize them. The confusion matrix below is the result of Decision Tree model with an accuracy rate of 68%.



accuracy: 68.00%

	true 1	true 2	true 3	class precision
pred. 1	0	0	0	0.00%
pred. 2	1	0	0	0.00%
pred. 3	3	4	17	70.83%
class recall	0.00%	0.00%	100.00%	

Figure 5.1: Decision tree performance- Target group Top 10 – 20 - 100

While comparing the other two classifications, a dataset which is divided into top 10 – 30 – 100 are likely to be easily predicted song performance than the top 10 – 50 –

100 songs. Beside higher accuracy level, the precision and recall rate also pointed out that songs in that group are equally divided and have some certain patterns within each subgroup that made prediction much easier. The confusion matrix below is the result of Decision Tree model of class 2: Top 10 – 30 – 100.

accuracy: 80.77%

	true 1	true 2	true 3	class precision
pred. 1	3	1	0	75.00%
pred. 2	1	6	2	66.67%
pred. 3	0	1	12	92.31%
class recall	75.00%	75.00%	85.71%	

Figure 5.2 Decision tree performance – Target group Top 10 – 30 - 100

All in all, target group as Top 10 – 30 – 100 was finally chosen to proceed next steps.

Correlation matrix to evaluate the contribution of each predictor to models

The interplays of all variables are indeed useful to build a predictive model. However, there is no certainty to assert the contribution of each variable to the whole predictive model. Therefore, correlations were calculated in order to measure the degree of associations between variables thanks to built-in Pearson Correlation tool in Rapid Miner.

Attributes	**Total tweets**	**No. of negative tweets**	**No. of positive tweets**	**Avg. negative score**	**Avg. positive score**	**No. of tweets abt artists**	**Weeks chart**	**Song ranking**
Total tweets	1	0.89	0.689	0.365	-0.063	0.494	-0.002	-0.372
No. of negative tweets	0.89	1	0.387	0.466	-0.049	0.344	-0.047	-0.222
No. of positive tweets	0.689	0.387	1	0.101	0.159	0.47	-0.07	-0.488
Avg. negative score	0.365	0.466	0.101	1	0.036	0.187	-0.195	-0.194
Avg. positive score	-0.063	-0.049	0.1509	0.036	1	-0.011	-0.051	0.006

No. of tweets abt artists	0.494	0.344	0.47	0.187	-0.011	1	0.198	-0.314
Weeks chart	-0.002	-0.047	-0.07	-0.195	-0.051	0.198	1	-0.036
Song ranking	-0.372	-0.222	-0.488	-0.194	0.006	-0.314	-0.036	1

Table 5.2: Pearson Correlation Matrix

To start with, how long a song has been in the chart has little or no correlation with the rest variables. Notably, it was also found to be barely correlated with target variable – song rankings. Meanwhile, surprisingly number of tweets about artists is relatively correlated with other variables. Initially, the author thought that number of tweets about artists has no contribution to predictive models. The data collection itself created a concern about the relevance of this variable. Tweets about artists are collected by directly crawling tweets that mentioned artists official Twitter account which actually made the data cover tweets not only related to artist and songs but also their private life, scandals or other random tweets about them. Besides, some artists do not have an official account that can also be a valid reason to question the validity of this variable. However, as can be seen from confusion matrix, total number of tweets about artists is relatively correlated with total tweets about songs r = 0.494. This might be interpreted that an increase in the number of tweets about songs leads to an increase in tweets about artists and vice versa. Between song ranking and independent variables, multiple correlations are also observed. To note, song performance is ranked from 1 to 3 in which 1 is the ranking of top 10 popular songs, 2 covers top 30 songs and 3 is to illustrated song ranked 31 – 100. Thus the negative correlation with target group as seen in confusion matrix actually turned out as a positive correlation with the song performance in reality. The findings showed that number of positive tweets seem to be the most critical predictor variable with r = 0.488, following by the number of tweets about artist and number of tweets about song with r = 0.372 and r = 0.314 respectively.

Compare the performance of combined data and single data source

Previous researches by Kim (2014) and Zangerle (2016) found that an additional variable which is number of weeks a song was in the chart increased accuracy of predictive model. This dissertation also wants to find out the predictive importance

of this variable under the restriction of time and data. Result of prediction was presented in table 2 below:

	Decision Tree	Neural Network	kNN
Twitter data	73.08	53.85	61.54
Combined data	80.77	61.54	61.54

Table 5.3 Accuracy rate of prediction models

The result of Pearson Correlation above showed that how long a song was in the chart has little correlation with the song rankings. However, predictive models with this additional variable actually outperform model built on solely Twitter data, which is in line with findings of earlier studies of Kim and Zangerle.

5.2. Discussion

The results from data analysis reveal a clear answer to the main research question of this study. It is found that the combination of sentiment-related variables, volume-related variables, and weeks on chart is able to predict song performances demonstrated in Billboard Hot 100 chart with 80.77% accuracy. This discussion will further analyze the results, and two main sub-questions are also going to be presented below.

Firstly, different classifications of song rankings were tested on predictive performance, and the classes of top 10 – 30 – 100 was chosen eventually. This is actually rather an empirical result than theory-based finding. Both the accuracy rate and confusion matrix were taken into account to find which is the best way to rank songs. Notwithstanding top 10 and non-top 10 classification provides the highest predictive result of 92% in the decision tree model, it was not selected to rank songs performance. Firstly the imbalanced class between top 10 and non-top 10 songs results in a high accuracy rate of predictive models because it is quite easy for classifier to tell the differences between hit and non-hit songs based on characteristics of variables set. Secondly, this binary target group was not the initial intention of this research. It was added to the modeling stage merely to compare the prediction result between binary target model with multiclass models. Imbalanced dataset once again faced against the classification of top 10 – 20 – 100 where 6 songs

belonged to top 10, 5 songs ranked top 20 whilst 22 songs ranked within a range of 21st to 100th. Class 3 of target group with top 10 – 30 – 100 was chosen for its superior prediction performance among all classification. Additionally, as can be seen in Figure 5.1 and 5.2, the weakness of model lies on the middle group where there is no clear split between this group and the top and the last group. The middle-group songs share the same characteristics with the other groups which consequently creates difficulty in classification. Nevetherless, it is importantly noted that these findings are restricted to this study only due to limited data availability and time constraints. Larger dataset or longitudinal studies might produce totally different results.

Subsequently, after choosing the most suitable way to rank songs, this study has presented the results of main research question as well as first sub-question with regard to performance of prediction with solely Twitter data. As can be seen from Table 5.3, all models produced better performance when combine existed Twitter data with an additional feature called total weeks a song has already been in the chart. This is in line with previous research by Kim et al. 2014 in which she found that the multivariate predictive models of combine data can predict the future success of a track accurately.

Although it has to be careful when comparing results of undertaken research and previous studies since they all have approached the problem differently and made different assumptions, though given no reason behind the contribution of this variable to the model stated by Kim, the author suggested an explanation as follow. The popularity of songs witnessed some slight changes or no change within two weeks. At the same time, several songs experienced a substantial change in their position in the chart. However, it is not sufficient to conclude songs performance by solely based on the ranking itself. Thus, how long a song has already been in the chart provides a better look into performance of song. “You need to calm down” dropped from 2nd to 13th right in the first releasing week, denoting that it did not perform so well and was not typically favored by music listeners like others song of Taylor Swift. On the other hand, the fact that the song "When the party is over" was already in the chart for 35 weeks while still holding the 45th position turns out to be an impressive performance even though in 36th week, the song was crossed out of Top 100 song chart. Or the song “The Gitup” jumped a big step from 51st in the first week even though it was in the list for only two weeks up until then to 29th in next week. It can be asserted that this song is quite popular among music lover community. (*Note: A challenge called #thegitupchallenge was initiated from this song during that time which could be a part of the reason why this song was typically favored than first weeks)* To

sum up, this additional variable not only increase the accuracy of predictive models but also could be considered as an explanatory variable to the song ranking itself.

Finally, the contribution of each predictor to the song rankings was demonstrated by Pearson Correlation matrix in Table 5.2. The results showed that number of positive tweets has the strongest correlation with song rankings than other variables (r = 0.488). Previous studies also found that volume-related variable, which is number of tweets about artists and songs make a significant contribution to the high accuracy predictive performance. It can be interpreted that well-known and high ranking songs are likely to receive more positive tweets from Twitter users than less popular songs. This is totally understandable in a real-life context where music lovers as the author interact more with popular songs than less favorite ones. However, one should not make a mistake between number of positive tweets and average positive score of a song. The average positive score, on the other hand, is barely correlated with any of the variables, including target variable. Thus, a higher positive score does not necessarily associate with a higher song performance. Conversely, trendy songs are even more prone to receive two-sided positive and negative reviews.

It is noted that number of positive tweets is not a purely volume-related variable, but it is a mixture of both volume and sentiment-related variables. Thus, this finding adds a valuable point to the current literature, which has not explored the addition of sentiment-related variables yet. It can be concluded that Twitter sentiment-related variables are valuable in building a predictive model to forecast song performance. Additionally, another volume-related variable which is total number of tweets related to songs were found as the second highest correlation variable with song rankings (r =0.372). The contribution of volume-related variables to the music industry have been found mostly in the investigation about the power of social media in forecasting album sales such as Dewand and Ramaprasad (2009), Cui et al. (2012), Dhar and Chang (2007). In a specific topic of this undertaken research, Kim (2014) and Zangerle (2016) also proved the importance of this factor in forecasting song ranking. Another variable which is slightly correlated with target variable which is number of tweets about artist (r = 0.314); nevertheless, the relationship is not as strong as number of positive tweets. Besides, Figure 5.3 below is the output of decision tree model which is built upon the song ranking classification of top 10 – 30 – 100 in which number of positive tweets also plays an essential role as a root of the tree splitting data

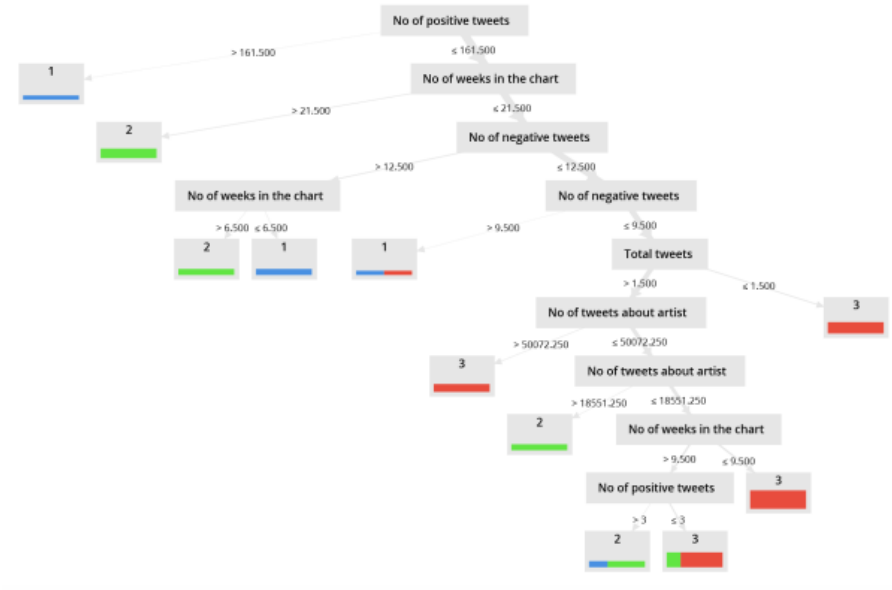


Figure 5.3: Decision Tree output with target group Top 10 – 30 - 100

Chapter VI: CONCLUSION & RECOMMENDATIONS

This chapter will discuss the overall conclusions of this study and recommendations for further researches. Hence, this section starts with a brief overview of the main goal of this study and how it was achieved, following the summary of main findings. These key findings, along with some limitations, subsequently become the premises to formulate recommendations of future researches.

6.1. Conclusion

The primary goal of this research is to explore the effect of online chatter in forecasting song performance. Unlike previous researches of the same topic which are longitudinal studies, this research faced difficulty in time restrictions and limited data availability, leading to the decision of following cross-section studies. Besides, this is the first study to investigate the contribution of sentiment-related variables in the music industry. Given all that conditions, there was no certainty in the research outcomes; therefore, research is following an exploratory approach to "seek new insights and assess the problem in a new perspective" (Saunders, Lewis and Thornhill, 2009). Twitter data, which includes volume-related and sentiment-related variables, were crawled directly from Twitter by Python with streaming APIs key is projected to song position in Billboard Hot 100 chart. Data analysis was performed in Rapid Miner by using different supervised classification algorithms to find the best performing model that can be utilized further by post studies.

Key findings of this research were found to be helpful not only in answering main research questions and two sub-questions but also helpful in gaining new insights into the music industry from a different perspective and adding values to the literature review by its new method of data collection. The results of this study suggested that Twitter data can be exploited as a predictor of song performance. But models perform better when incorporating an additional variable of weeks a song was in Billboard chart which can be considered as an explanatory variable for the target group song ranking to have a better picture of how the song actually performed over time. The finding is consistent with earlier studies by discovering the contribution of volume-related variables to the whole predictive models. Besides, the results of undertaken research has pointed out the contribution of sentiment-related

variables from Twitter to the music industry. Another highlight of empirical result is that number of positive tweets as both volume-related and sentiment-related variable is the most powerful predictor to song ranking charts. This is an important observation of the research as to the best of author's knowledge, none of previous studies investigated the contribution of sentiment analysis from social media in music industry. It can be concluded that sentiment-related variables is useful as a proxy of song rankings. But it is not possible to generalize how effective it is. Since this undertaken research is performed under data and time restrictions, empirical results can be different when employed in larger dataset. Besides, notwithstanding all Pearson correlation coefficients were relatively low under the average baseline of 50%, multivariate models with the combination of all independent variables performed pretty well in predicting song rankings with an accuracy rate of 80.77% (as illustrated in Table 5.3). Then, after evaluating the prediction results of three different classification algorithms, decision tree prediction model which is known for its easy to extract display rule yet high classification precision (Tran, 2019) was chosen for its superior performance than others.

6.2. Limitations and directions for future researches.

This study has faced several challenges, and one of them is limited data availability. Given the tremendous amount of Twitter data, this undertaken research only exploits a small partition of it. Firstly, as the primary data of this research are tweets that are related to songs and artists, the task of choosing appropriate song title and artist name hinders number of data collected. While collecting tweets about artists are simplified by mention function (@ + artist name), song titles have different variants, and there is no correct way to formulate the hashtags. Besides, even though using mention + artists official account to collect tweets related to artists was faster, and more conveniently, the habit of Twitter users to include hashtag artist name in their tweets results in a quite loss of tweets. More importantly, many song titles are too familiar, which might lead to an enormous amount of irrelevant data therefore from 100 songs in Hot 100 chart, only 37 songs were selected to form the final dataset. Besides, handling unstructured Twitter data is another limitation of this study. It is undoubtedly that social media data which contains various components such as grammatical error, slangs, abbreviations, and misspellings require an effort to turn those data into readable and easy-to-interpret data. But when it comes to tweets about music, crawled tweets are even noisier and messier than other types of text. People are likely to insert URL (hyperlink) of songs, emojis to express their opinions,

and both hashtags and mentions to call out name of songs and artists. Especially when dealing with such noisy tweets, the matching keywords is a crucial task as the quality of data resulting from this step significantly influences the quality of recommendation. In this study, there were some typo mistakes in mentioning artist's official Twitter account, for example, @kattyperry instead of @kateperry, resulting in the omission of some songs from the list merely due to this mistakes.

Last but not least, like most research projects undertaken for academic purposes are necessarily time-constrained (Saunders, Lewis and Thornhill, 2009) and this paper is not an exception. Limited data availability creates difficult condition for classifier model to specify the differences between song rankings each other. Hence, lowest class precision was found when predicting the middle group songs, which shares the same trait with the first group and last group. The lack of time and data is also the main reason for the decision of ranking songs as subgroup such as Top 10 – 30 – 100 instead of predicting exact song position in the chart. Besides, the results of this study seem to be exclusive to music industry due to the method of choosing keywords (artist name and song title). Thus different types of keywords would be needed when exploring the power of Twitter in other industry.

Based on key findings and limitation presented, there are several suggestions for further developments in the future. Firstly, based on given results and limitations of this study, the author suggests future researches pursue longitudinal study, which is the research carried out over a long period of time. The study of Kim crawled Twitter data over the course of 10 weeks. However, Zangerle (2016) asserted that 10-week was still a limited analysis period that did not capture all information needed. Zangerle's study was thus based on data collected constantly in 2 years, covering a time frame substantially longer than previous research, utilizing time series to perform analysis (Zagerle et al., 2016). Longer previous time also associates with a larger number of data collected. Therefore, it will not allow researchers to have a better look at the problems and discover time-related issues but also provide more time for machine learning algorithms….Secondly, according to Wikipedia, the Billboard chart reflects the US music market only. However, it was still chosen for it is one of the most influential indicators for songs popularity (Zangerle, 2016). Future researches could refine results by using other data source for song ranking charts beside Billboard such as Youtube chart, Spotify chart, Itunes chart, which are reliable and international-scale charts.

Besides, the topic of undertaken research, which is leveraging Twitter to predict song ranking performance could be useful to some other extent by following ways. Firstly, Twitter, Instagram, and Facebook are three main social networking sites where people are able to freely express their opinions, producing a huge amount of data about users behaviors. Thus these sites could potentially be the primary data source for future researches. Besides, if song performance was quantized into song rankings chart, movie performance was also illustrated in movie ranking charts (such as IMDb, Rottentomatoes). While there are numerous available studies exploring the relationship between movie offices and online chatter, to the best of author's knowledge, there is no prior research which aims to investigate the association between movie ranking and social networking sites.

Lastly, the use of sentiment-related variables could be extended in further researches. A larger amount of Twitter data would, therefore, reduce sample bias of users' attitudes (15) Probably the way to calculate the average positive and negative score in this research needs further thorough research before putting them in usage in predictive model. Besides, other variables could also be collected from Twitter such as number of retweets, location of the tweets posted or number of followers an author has on Twitter which was exploited by a research of Rui (2011) in order to predict movie box office revenues. The contribution of each variable will vary from different research topic to others. None of the previous researches which discover Twitter data and song ranking has employed any of those variables above; hence, the author does not have the certainty to suggest of those. Future researches could consider them to formulate independent variables set in order to build a better predictive model to forecast song performance.

APPENDIX A - PYTHON CODE

A small part of Python code to crawl Twitter data via Streaming APIs

```python
from tweepy import OAuthHandler
from tweepy import Stream
from tweepy.streaming import StreamListener

access_token = "457694877-DQCTsl8Fa0vG6EIo9oFJDgXdeJe2tIkIDOq6r9De"
access_secret = "B3R6EZSJGON4cFMQfsz2kFALVobJeXPNqW0qU3Fx5FPKO"
consumer_key = "1GzHn7pCUWelJADA0y13HnqEL"
consumer_secret = "phIjiv8pfC93peg7Q4LWEmHC7ErCz8nh5ca5tK5Ca3hsDLN6Gy"

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)


class MyListener(StreamListener):
    def on_data(self, data):
        try:
            with open('./data/1213.json', 'a') as f:
                f.write(data)
                print('>', end='', flush=True)
                return True
        except BaseException as e:
            print("Error on_data: %s" % str(e))
        return True

    def on_error(self, status):
        print(status)
        return True

songAndArtist = ['#youneedtocalmdown', '@taylorswift',
                 '#badguy', '@billieeilish',
                 '#idontcare', '@edsheeran', '@justinbieber',
                 '#moneyinthegrave', '@rickross', '@drake',
                 '#suge', '@dababydababy',
                 '#noguidance', '@chrisbrown', '@drake',
                 '#dancingwithastranger', '@samsmith', '@normani',
                 '#ificanthaveyou', '@shawnmendes',
                 '#truthhurts', '@lizzo',
                 '#withoutme', '@halsey',
                 '#7rings', '@arianagrande',
                 '#godscountry', '@blakeshelton',
                 '#whiskeyglasses', '@morganwallen',
                 '#heylookmaimadeit', '@panicatthedisco',
                 '#thelondon', '@youngthug', '@jcodenc', '@trvisXX',
                 '#concalma', '@daddyyankee', '@katyperry',
                 '#beerneverbrokemyheart', '@lukecombs',
                 '#neverreallyover', '@kattyperry',
                 '#earfquake', '@tylethecreator',
                 '#lookwhatgodgaveher', '@thomasrhett',
                 '#whenthepartysover', '@billieeilish',
                 '#crossme', '@edsheeran',
                 '#goloko', '@YG', '@Tyga','@JonZ',
                 '#thegitup', '@blancobrown',
                 '#someoneyouloved', '@lewiscapaldi',
                 '#walkmehome', '@pink',
                 '#knockinboots', '@lukebryan',
                 '#girlsneedlove', '@iamsummerwalker', '@drake',
                 '#alltomyself', '@danandshay',
                 '#shottaflow', '@nlechoppa1',
                 '#callaita', '@imbadbunny',
                 '#rearviewtown', '@jasonaldean',
```

APPENDIX B - FINAL DATASET

Songs	Total tweets	No of neg tweets	No of pos tweets	Avg neg score	Avg pos score	No of tweets abt artist	Weeks in the chart	Ranking
#youneedtocalmdown	302	17	217	0.016470199	0.127099338	29708	1	2
#badguy	221	64	47	0.089298643	0.058217195	86492	12	3
#idontcare	190	14	101	0.021526316	0.1526	65373.5	6	5
#moneyinthegrave	102	22	30	0.056843137	0.086039216	2024	1	7
#suge	46	11	11	0.044521739	0.0705	70	12	9
#noguidance	65	21	12	0.056046154	0.037907692	22173	2	10
#dancingwithastranger	9	0	1	0	0.032555556	4163	23	12
#ificanthaveyou	121	13	16	0.018553719	0.02653719	16897	7	13
#truthhurts	1166	794	106	0.105962264	0.089554031	45627	7	14
#withoutme	24	1	12	0.013458333	0.100291667	117658	37	16
#7rings	387	23	127	0.013850129	0.076511628	3710	22	19
#godscountry	94	8	34	0.017468085	0.099808511	5135	12	21
#whiskeyglasses	14	0	2	0	0.025071429	15	16	22
#heylookmaimadeit	6	0	4	0.048	0.238333333	67	10	24
#thelondon	11	3	3	0.054090909	0.074545455	24585	4	26
#concalma	59	1	12	0.003322034	0.051135593	20930	20	27
#beerneverbrokemyheart	24	1	13	0.022833333	0.174875	23001	7	28
#neverreallyover	89	6	25	0.009292135	0.087831461	41294	3	29
#lookwhatgodgaveher	1	0	0	0	0	15	16	37
#whenthepartysover	19	3	5	0.029	0.050421053	86492	32	45
#crossme	10	3	5	0.0772	0.1123	66572	4	46
#goloko	22	1	4	0.011136364	0.042318182	2629.333333	7	49
#thegitup	180	8	137	0.005766667	0.152983333	450	2	51
#someoneyouloved	37	1	17	0.017972973	0.140162162	1961	6	54
#walkmehome	2	0	1	0	0.125	1956	16	55
#alltomyself	6	1	1	0.025333333	0.087666667	17	7	66
#callaita	60	12	5	0.053183333	0.03885	2	9	68
#rearviewtown	2	0	1	0.058	0.0965	5	4	69
#sanguineparadise	8	0	7	0.020625	0.337625	116	10	78
#herewithme	19	2	2	0.018052632	0.024368421	12517.5	15	79
#otrotrago	27	3	3	0.018407407	0.022481481	213	2	81
#callyoumine	15	0	5	0.010333333	0.106466667	1630	3	82

#raisedoncountry	27	2	20	0.027259259	0.224518519	35	3	83
#loveaint	0	0	0	0	0	4	10	85
#baccatitagain	5	0	1	0	0.0462	1656.5	4	87
#terobare	1	0	1	0.073	0.267	72	6	94
#racksinthemiddle	6	1	0	0.036833333	0.033666667	211.3333333	11	99
#theonesthatdidntmakeitbackhome	1	0	1	0	0.185	15	1	100
#youneedtocalmdown	134	11	75	0.018619403	0.100149254	13674	2	13
#badguy	1012	435	173	0.100578063	0.042959486	290218	13	3
#idontcare	399	23	289	0.016348371	0.228716792	174928	7	5
#moneyinthegrave	48	15	13	0.075645833	0.095854167	2077	2	8
#suge	27	2	6	0.027925926	0.04237037	49	13	7
#noguidance	107	13	40	0.047074766	0.110261682	19869.5	3	9
#dancingwithastranger	14	0	4	0	0.0445	10776.5	24	15
#ificanthaveyou	123	15	6	0.019097561	0.012479675	9053	8	14
#truthhurts	776	489	110	0.105739691	0.096265464	37611	8	11
#withoutme	23	0	14	0.004086957	0.198521739	93437	38	17
#7rings	428	19	150	0.008698598	0.076813084	10156	23	23
#godscountry	85	5	36	0.014541176	0.122258824	6180	13	18
#whiskeyglasses	13	0	4	0	0.087153846	10	17	28
#heylookmaimadeit	2	0	1	0	0.149	131	11	24
#thelondon	5	1	2	0.0362	0.1288	24870.5	5	30
#concalma	234	0	6	0	0.00674359	58850.5	21	37
#beerneverbrokemyheart	36	3	17	0.019305556	0.126833333	8062	8	33
#neverreallyover	137	5	21	0.009138686	0.046058394	117324	4	27
#lookwhatgodgaveher	2	1	1	0.1695	0.08	23	17	54
#crossme	23	1	8	0.02473913	0.080695652	174928	5	39
#goloko	25	1	10	0.01152	0.09336	4073.666667	8	52
#thegitup	106	9	67	0.015490566	0.132801887	851	3	29
#someoneyouloved	265	0	15	0.001332075	0.013158491	167	7	53
#walkmehome	7	0	1	0.008428571	0.042857143	1058	17	62
#knockinboots	33	0	27	0	0.158393939	10	9	61
#girlsneedlove	1	0	1	0	0.412	2049	18	74
#alltomyself	6	1	2	0.0335	0.114166667	16	8	67
#callaita	25	6	2	0.07336	0.05228	4	3	68
#rearviewtown	1	0	1	0.116	0.193	1	5	71
#sanguineparadise	3	0	0	0	0	40	11	89
#herewithme	10	0	1	0	0.0217	12241	16	87

#otrotrago	26	7	0	0.050961538	0	303	3	83
#callyoumine	18	2	3	0.032944444	0.039055556	608.5	4	73
#raisedoncountry	38	2	27	0.014026316	0.211578947	87	5	81
#loveaint	2	0	1	0	0.155	8	11	90
#baccatitagain	4	0	1	0	0.04025	1940	5	78
#theonesthatdidntmakeitbackhome	0	0	0	0	0	20	2	93

BIBLIOGRAPHY

Aggarwal, C. and Zhai, C. (2012). Mining text data. New York: Springer.

Allahyari, Mehdi & Pouriyeh, Seyedamin & Assefi, Mehdi & Safaei, Saied & Trippe, Elizabeth & Gutierrez, Juan & Kochut, Krys. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.

Asur, Sitaram & Huberman, Bernardo. (2010). Predicting the Future with Social Media. Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010. 1. 10.1109/WI-IAT.2010.63.

Bali, R., Sarkar, D. and Sharma, T. (2017). Learning social media analytics with R. 1st ed. Birmingham: Packt Publishing Ltd., p.7.

Billboard. (2017). Billboard Charts to Adjust Streaming Weighting in 2018. [online] Available at: https://www.billboard.com/articles/business/8006673/billboard-charts-adjust-streaming-weighting-2018

Billboard. (2018). Billboard Finalizes Changes to How Streams Are Weighted for Billboard Hot 100 & Billboard 200. [online] Available at: https://www.billboard.com/articles/news/8427967/billboard-changes-streaming-weighting-hot-100-billboard-200

Billboard. (n.d.). Top 100 Songs | Billboard Hot 100 Chart. [online] Available at: https://www.billboard.com/charts/hot-100

Bird, S., Klein, E. and Loper, E. (2009). Natural Language Processing with Python. 1st ed. O'Reilly Media.

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1).

Cement, J. (2019). Most followed accounts on twitter 2019 | Statista. [online] Available at: https://www.statista.com/statistics/273172/twitter-accounts-with-the-most-followers-worldwide/

Chong, Alain & Ngai, Eric & Ch'ng, Eugene & Li, Boying & Lee, Filbert. (2015). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. International Journal of Operations & Production Management.

Collins, S. (2013). Spotify: merging music with social media. [online] The Conversation. Available at: https://theconversation.com/spotify-merging-music-with-social-media-18401

Crupnick, R. (2018). Music Scores A Gold Record on The Social Media Charts | MusicWatch Inc.. [online] Musicwatchinc.com. Available at: https://www.musicwatchinc.com/blog/music-scores-a-gold-record-on-the-social-media-charts/

Cui, Geng & Lui, Hon-Kwong & Guo, Xiaoning. (2012). The Effect of Online Consumer Reviews on New Product Sales. International Journal of Electronic Commerce. 17. 39-58. 10.2307/41739503.

Dewan, Sanjeev & Ramprasad, Jui. (2009). Chicken and Egg? Interplay between Music Blog Buzz and album Sales.. 87.

Dhar III, Vasant & A. Chang, Elaine. (2007). Does Chatter Matter? The Impact of User-Generated Content on Music Sales. Journal of Interactive Marketing. 23. 10.2139/ssrn.1113536.

E. Tunggawan and Y. E. Soelistio, "And the winner is …: Bayesian Twitter-based prediction on 2016 U.S. presidential election," 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, 2016, pp. 33-37.

En.wikipedia.org. (n.d.). Application programming interface. [online] Available at: https://en.wikipedia.org/wiki/Application_programming_interface

En.wikipedia.org. (n.d.). Billboard Hot 100. [online] Available at: https://en.wikipedia.org/wiki/Billboard_Hot_100

En.wikipedia.org. (n.d.). Billboard charts. [online] Available at: https://en.wikipedia.org/wiki/Billboard_charts

En.wikipedia.org. (n.d.). Rapid Miner. [online] Available at: https://en.wikipedia.org/wiki/RapidMiner

En.wikipedia.org. (n.d.). Record chart. [online] Available at: https://en.wikipedia.org/wiki/Record_chart

Fan, Weiguo & Gordon, Michael. (2014). The Power of Social Media Analytics. Communications of the ACM. 57. 74-81. 10.1145/2602574

Franklin, K. (2013). Social media is revolutionising the music industry. Retrieved May 11, 2015, from https://www.brandwatch.com/2013/08/social-media-the-music-industry

Gupta, Bhumika & Negi, Monika & Vishwakarma, Kanika & Rawat, Goldi & Badhani, Priyanka. (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python. International Journal of Computer Applications. 165. 29-34. 10.5120/ijca2017914022.

Gupta, Vishal & Lehal, Gurpreet. (2009). A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. 1. 10.4304/jetwi.1.1.60-76

H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu and B. Liu, "Predicting Flu Trends using Twitter data," 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, 2011, pp. 702-707

Help.twitter.com (n.d.). How to use hashtags. [online]. Available at

Help.twitter.com. (n.d.). About Twitter's APIs. [online] Available at: https://help.twitter.com/en/rules-and-policies/twitter-api

Help.twitter.com. (n.d.). About replies and mentions. [online] Available at: https://help.twitter.com/en/using-twitter/mentions-and-replies

Hutchinson, A. (2016). Here's Why Twitter is so Important, to Everyone. [online] Social Media Today. Available at: https://www.socialmediatoday.com/social-networks/heres-why-twitter-so-important-everyone

IFPI (2018). Global Music Report 2018. [online] IFPI. Available at: https://www.ifpi.org/downloads/GMR2018.pdf

Jha, A. (2012). Web Crawling: Data Scraping vs. Data Crawling | | PromptCloud. [online] Promptcloud.com. Available at: https://www.promptcloud.com/blog/data-scraping-vs-data-crawling/

Kao, A. and Poteet, S. (2010). Natural Language Processing and Text Mining. London: Springer.

Kim, Y., Suh, B., & Lee, K. (2014). #nowplaying the Future Billboard: Mining Music Listening Behaviors of Twitter Users for Hit Song Prediction. SoMeRA@SIGIR.

Marketing Charts. (2017). We Listen to Music For More Than 4 1/2 Hours A Day, Nielsen Says - Marketing Charts. [online] Available at: https://www.marketingcharts.com/industries/media-and-entertainment-81082

Molla, R. and Kafka, P. (2018). Here's why the music industry is celebrating again — and here's why the music industry is still in mourning. [online] Vox. Available at: https://www.vox.com/2018/1/27/16933704/music-streaming-charts-2010-1999-spotify-sales-grammys-cardi-b

Mulligan, M. (2019). Facebook could be the future of social music but isn't yet. [Blog] Music Industry Blog. Available at: https://www.hypebot.com/hypebot/2019/03/facebook-could-be-the-future-of-social-music-but-isnt-yet-mark-mulligan.html

Nabi, J. (2018). Machine Learning — Multiclass Classification with Imbalanced Dataset. [Blog] Towards Data Science. Available at: https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a

Nguyen H., Nguyen TN., Danny Dig, Nguyen S., Tran H., and Hilton M. Graph-based mining of in-the-wild, fine-grained, semantic code change patterns. In Proceedings of the 41st International Conference on Software Engineering (ACM/IEEE ICSE 2019). Montreal, Canada — May 25 - 31, 2019.

O'Dea, B., Wan, S., Batterham, P., Calear, A., Paris, C. and Christensen, H. (2015). Detecting suicidality on Twitter. Internet Interventions, 2(2).

Orabi A.H., Buddhitha P., Orabi M.H., Inkpen D., (n.d.). Deep Learning for Depression Detection of Twitter Users. [online] Available at: https://aclweb.org/anthology/W18-0609

Patel, P. (2018). Why Python is the most popular language used for Machine Learning - Medium. [online] Medium.com. Available at: https://medium.com/@UdacityINDIA/why-use-python-for-machine-learning-e4b0b4457a77

Provost, F. and Fawcett, T. (2013). Data science for business. Sebastopol, Calif.: O'Reilly.

Reiman, M., & Örnell, P. (2018). Predicting Hit Songs with Machine Learning (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-229705

Rui, Huaxia & Liu, Yizao & Whinston, Andrew. (2011). Whose and What Chatter Matters? The Impact of Tweets on Movie Sales. Decision Support Systems. 55. 10.2139/ssrn.1958068

Salloum, Said & Al-Emran, Mostafa & Abdel Monem, Azza & Shaalan, Khaled. (2017). A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives. Advances in Science, Technology and Engineering Systems Journal. 2. 127-133. 10.25046/aj020115.

Saunders, M., Lewis, P. and Thornhill, A. (2009). Research methods for business students. 5th ed.

Snowflake Software. (n.d.). REST APIs & Streaming APIs: Aviation Data Access for Everyone. [online] Available at: https://snowflakesoftware.com/news/rest-apis-streaming-apis/

Talib, Ramzan & Kashif, Muhammad & Ayesha, Shaeela & Fatima, Fakeeha. (2016). Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.071153.

Throckmorton, K. (n.d.). How Technology Has Changed How We Listen to Music. [online] Sutori.com. Available at: https://www.sutori.com/story/how-technology-has-changed-how-we-listen-to-music--7iyyK1aQqRuj8hbS7HQqKoj9

Tran H., Shcherbakov M. (2016) Detection and Prediction of users attitude based on real-time and batch sentiment analysis of facebook comments. In: Nguyen H.,

Snasvel V. (eds) Computational Social Networks. CSoNet 2016. Lecture Notes in Computer Science, vol 9795. Springer, Cham

Tran, H. A survey of machine learning and data mining techniques used in multimedia system. Sep. 2019

Tran H., Tran N., Nguyen S., Nguyen H., Nguyen TN. (2019). Recovering variable names for minified code with usage contexts. In Proceedings of the 41st International Conference on Software Engineering (ACM/IEEE ICSE 2019). Montreal, Canada — May 25 - 31, 2019.

Tran N., Tran H., Nguyen S., Nguyen H., Nguyen TN. (2019). Does BLEU score work for code migration? In Proceedings of the 27th International Conference on Program Comprehension (ACM/IEEE ICPC 2019). Montreal, Canada — May 25 - 31, 2019.

Vossen, R. (n.d.). Does Chatter Matter? Predicting Music Sales with Social Media. [online] Available at: https://www.basicthinking.de/blog/wp-content/uploads/2013/06/Does-Chatter-Matter.pdf

Wirth, R & Hipp, Jochen. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Woodford, C. (2019). How neural networks work - A simple introduction. [online] Explain that Stuff. Available at: https://www.explainthatstuff.com/introduction-to-neural-networks.html

Zangerle, E., Pichl, M., Hupfauf, B. and Specht, G. (2016). CAN MICROBLOGS PREDICT MUSIC CHARTS? AN ANALYSIS OF THE RELATIONSHIP BETWEEN #NOWPLAYING TWEETS AND MUSIC CHARTS. [online] Available at: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/039_Paper.pdf

Zimmer, Michael & Proferes, Nicholas. (2014). A Topology of Twitter Research: Disciplines, Methods, and Ethics. Aslib Journal of Information Management. 66. 10.1108/AJIM-09-2013-0083.