

Lyrics Matter

Using Lyrics to Solve Music Information Retrieval Tasks

by

Abhishek Singhi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2015

© Abhishek Singhi 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Music Information Retrieval (MIR) research tends to focus on audio features like melody and timbre of songs while largely ignoring lyrics. Lyrics and poetry adhere to a specific rhyme and meter structure which set them apart from prose. This structure could be exploited to obtain useful information, which can be used to solve Music Information Retrieval tasks. In this thesis we show the usefulness of lyrics in solving MIR tasks. For our first result, we show that the presence of lyrics has a variety of significant effects on how people perceive songs, though it is unable to significantly increase the agreement between Canadian and Chinese listeners about the mood of the song. We find that the mood assigned to a song is dependent on whether people listen to it, read the lyrics or both together. Our results suggests that music mood is so dependent on cultural and experiential context to make it difficult to claim it as a true concept. We also show that we can predict the genre of a document based on the adjective choices made by the authors. Using this approach, we show that adjectives more likely to be used in lyrics are more rhymable than those more likely to be used in poetry and are also able to successfully separate poetic lyricists like Bob Dylan from non-poetic lyricists like Bryan Adams. We then proceed to develop a hit song detection model using 31 rhyme, meter and syllable features and commonly used Machine Learning algorithms (Bayesian Network and SVM). We find that our lyrics features outperform audio features at separating hits and flops. Using the same features we can also detect songs which are likely to be shazamed heavily. Since most of the Shazam Hall of Fame songs are by upcoming artists, our advice to them is to write lyrically complicated songs with lots of complicated rhymes in order to rise above the “sonic wallpaper”, get noticed and shazamed, and become famous. We argue that complex rhyme and meter is a detectable property of lyrics that indicates quality songmaking and artisanship and allows artists to become successful.

Acknowledgements

I am grateful to everyone in the gym at CIF who has allowed me to work in, spotted me, helped me improve my form or taught me a new exercise. Some of my best times at Waterloo have been spent in the gym.

I would like to thank my advisor Prof Dan Brown for his guidance, encouragement and patience. He has always been open to new ideas, providing quick and honest feedback, and has given me the freedom to pursue them. His assistance and insight has been invaluable during my masters. I am also thankful to my readers Prof Charlie Clarke and Prof Daniel Vogel for taking the time to evaluate and provide feedback on my thesis.

I am exceedingly grateful to my uncle Jeetendra Falodia for helping me have a smooth transition from India to Canada. He has always been there for me, offering me guidance and encouragement when needed, and in general has been a pretty cool uncle. Thank you to my friends Idhayachandhiran Ilampooranan, Rahul Deshpande, Sandeep Vridhagiri and Dinesh Alapati for all the fun times we had, whether in Waterloo or during one of our many road trips.

Dedication

Kate, thanks for making me realize that I can.

Table of Contents

List of Tables	x
List of Figures	xiii
1 Introduction	1
1.1 Lyrics and Music Mood Perception	2
1.2 Lyrics vs Poetry	2
1.3 Hit Song Detection	3
1.4 How To Get Shazamed?	3
1.5 Summary	4
2 Background and Related Work	5
2.1 Lyrics and Neuroscience	5
2.2 Lyrics and Music Information Retrieval	6
3 On Cultural, Textual And Experiential Aspects Of Music Mood	8
3.1 Introduction	8
3.2 Related Work	10
3.2.1 What is Mood Detection?	10
3.2.2 Methods for Detecting Mood	10
3.2.3 Mood Tags	10

3.2.4	Music Mood Perception between different Cultures	11
3.3	Method	11
3.3.1	Data Set	11
3.3.2	Participants	13
3.3.3	Survey	13
3.4	Results	14
3.4.1	Lyrics and music mood perception between cultures	14
3.4.2	Stability across the three kinds of experiences	16
3.4.3	“Melancholy” lyrics	17
3.4.4	Genre and Mood	17
3.5	Conclusions: Does Music Mood Exist?	19
4	Are Poetry and Lyrics All That Different?	21
4.1	Introduction	21
4.1.1	Definitions and Synonyms	21
4.2	Related Work	23
4.3	Data Set	24
4.3.1	Articles	25
4.3.2	Lyrics	25
4.3.3	Poetry	25
4.3.4	Test Data	26
4.4	Method	26
4.4.1	Extracting Synonyms	27
4.4.2	Probability Distribution	28
4.4.3	Document classification algorithm	28
4.5	Results	29
4.5.1	Document Classification	30
4.5.2	Adjective Usage in Lyrics versus Poems	30

4.5.3	Poetic vs non-Poetic Lyricists	32
4.5.4	Concept representation in Lyrics vs Poetry	33
4.6	Conclusion	33
5	Can Song Lyrics Predict Hits?	36
5.1	Introduction	36
5.2	Related Work	37
5.3	Rhymes	39
5.3.1	Imperfect Rhymes	39
5.3.2	Internal Rhymes	40
5.4	Data Definition	41
5.5	Method	43
5.6	Results	46
5.6.1	The Broadest Definition of Flop	46
5.6.2	The “Flash in a Pan” Definition of Flop	50
5.6.3	The “Hit on One Chart” Definition of Flop	51
5.6.4	The “Not-Hit Single” Definition of Flop	51
5.7	What makes a song hit?	52
5.8	Conclusion: Lyrics Complexity and Craftmanship	53
6	Lyrics Complexity and Shazam	54
6.1	Introduction	54
6.2	Related Work	56
6.3	Data	56
6.4	Method	57
6.4.1	Lyrics Complexity	57
6.5	Shazam Users	58
6.6	Results	58

6.6.1	Separating the Shazam Hall of Fame songs from Hits using the lyrics features	59
6.6.2	Lyrics complexity and shazams	61
6.6.3	Shazam Hall of Fame and Chorus Complexity	65
6.7	Conclusion: Advice for musicians	67
7	Conclusion	69
	References	74

List of Tables

3.1	The mood clusters used in the study.	11
3.3	The list of names, artists and the year of release of songs which were used in the study.	13
3.4	The number of statistically-significantly similar responses between the different cultures for the three different ways they interact with the songs. Canadians refer to Canadians of both Chinese and non-Chinese origin.	16
3.5	Spearman’s rank correlation coefficient between the groups. The groups “only lyrics” and “only audio” identify when participants had access to only lyrics and audio respectively while “audio+lyrics refers to when they had access to both simultaneously.	16
3.6	The most commonly assigned mood clusters for each experimental context. Most songs are assigned to the third mood cluster when participants are shown only the lyrics.	17
3.7	Entropy values for the mood assignment of rock songs, for the three different categories of interactions with a song. An entropy value of 0 indicates that everyone agreed on the mood of the song.	19
3.8	Entropy values for hip-hop/ rap songs for the three different categories for each of the three ways people interact with a song. An entropy value of 0 indicates that everyone agreed on the mood of the song.	19
4.1	Fifteen of the singers and poets in our data set.	24
4.2	The list of poetic lyricists in our test set.	26
4.3	The probability distributions for the adjectives in the lyrics snippet in Figure 4.1. The overall classification for the snippet is that it is lyrics.	29

4.4	The confusion matrix for document classification. Many lyrics are categorized as poems, and many poems as articles.	31
4.5	Statistical values for the number of words an adjective rhymes with. The two-tailed P value is less than 0.0001.	31
4.6	Statistical values for the semantic orientation of adjectives used in lyrics and poetry. The two-tailed P value equals 0.8072.	31
4.7	Percentage of misclassified lyrics as poetry for poetic lyricists.	32
4.8	Percentage of misclassified lyrics as poetry for non-poetic lyricists.	33
4.9	For twenty different concepts, we compare adjectives which are more likely to be used in lyrics rather than poetry and vice versa.	34
5.1	The number of hits and flops in our data set using the four definitions of flops.	41
5.2	Fifteen of the artists in our data set.	42
5.4	The list of lyric features used by our algorithm. Singles per rhyme is the only feature that argues against lyrical complexity and is more relevant in Chapter 6.	45
5.5	Stress pattern in different meters.	46
5.6	The results we obtain using a weighted-cost SVM. The weights are chosen to keep the recall close to 50%.	47
5.7	The results we obtain using a Bayesian network.	47
5.8	Lyrics features outperform the audio features in discerning hits from flops.	49
5.9	Surprisingly, the naïve Bayesian network gives us better result than weighted-cost SVM when using both audio and lyrics features.	49
5.10	We see a noticeable improvement in performance with lyrics longer than 50 lines, which are more accurate than the shorter ones. Compare to Table 5.7	50
5.11	The most important features and their values for hit detection and the percentage of hits and flops falling in that range.	52
5.12	The outcome of our algorithm on hits in our data set using the first definition of flops for four popular artists.	52
6.1	The number of Shazam users and the number of smartphones sold worldwide.	58

6.2	The results for separating Shazam Hall of Fame songs from hits using a weighted-cost SVM	59
6.3	The results for separating Shazam Hall of Fame songs from hits using a Bayesian Network. The SVM does a better job at predicting hits, though the Bayes Net classifier is more stringent	60
6.4	The top five most lyrically-complicated Hall of Fame songs.	65
6.5	The table lists Hall of Fame songs having the chorus as the lyrically most complex and least complex part of the song.	67

List of Figures

4.1	The bold-faced words are the adjectives our algorithm takes into account while classifying a document, which in this case in a snippet of lyrics by the Backstreet Boys.	22
5.1	The receiver operating characteristic curve obtained when using the first and second definition of flops respectively. The areas under the ROC curves are 0.688 and 0.573 respectively.	48
5.2	The ROC curve obtained when using the first definition of flop and a weighted-cost SVM. The black and the red curves are obtained using the lyrics and audio features respectively. The AUC using the lyrics and audio features is 0.692 and 0.572 respectively.	50
6.1	An overlapped histogram of the lyrics complexity of the hits and Hall of Fame songs. The red coloured part is the histogram for the Hall of Fame songs while the histogram for hits is coloured green. Despite having far more hits than Hall of Fame songs, the majority of songs with lyric complexity measure greater than 8 are from the Hall of Fame. The overlapped portion is coloured dark green.	62
6.2	A scatter plot showing the relationship between the number of shazams and lyrics complexity for the songs in the 2013 Billboard Year-End Hot 100 singles chart (Correlation coefficient = 0.3986). The curve of best fit has R-squared value of 0.1976. Lyrically complex songs tends to be shazamed more.	63

6.3	A scatter plot showing the relationship between the number of shazams and lyrics complexity for the songs in the 2008 Billboard Year-End Hot 100 singles chart (Correlation coefficient = -0.1757). The curve of best fit has R-squared value of 0.0564. In the early years, the relationship between shazams and lyric complexity is less clear.	64
6.4	A scatter plot showing the relationship between the number of shazams and the 2014 Billboard Year-End Hot 100 rank for the songs which made it to the year end chart. “Blurred Lines” was the #2 song in 2013, which explains its outlier status for 2014.	66

Chapter 1

Introduction

Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music, focusing on the use of audio signals, lyrics and metadata of songs. Typical MIR applications include recommender systems, automatic music transcriptions and automatic categorization (into mood and genre, for example). Much of the research in MIR tends to focus on the audio signal of the music, specifically melodic or timbral features or on meta-tagged data of songs [56]. Some MIR tasks have used text features associated with the semantics and grammar of the words along with the audio features [45]. Lyrics features like rhyme and meter, that provide structure to the lyrics and separate it from prose, have largely been ignored.

In this thesis, we show the usefulness of lyrics features in solving some important MIR tasks. Lyrics contain much of the emotional content of a typical popular song [2] and can contribute to the memorizability of songs if they have catchy rhyme [89]. Behavioral and neuropsychological research has shown that individuals process lyrics and tune separately while listening to songs [37]. We show that the presence of lyrics does influence the way a song is perceived by the listener. We also show that different genres of writing use different adjectives for the same concept and using this observation we are able to separate lyrics from poetry. Using 31 rhyme, meter and syllable features we are able to separate hits from flops surprisingly well. We show that lyrics features outperform audio features at hit detection. We also find that the presence of lots of rhymes, in particular imperfect and internal rhymes, makes it more likely that a song will end up being a hit. We also find that the songs in the Shazam Hall of Fame [67] are lyrically more complicated and in particular have more complex rhymes than hits that are less shazamed.

1.1 Lyrics and Music Mood Perception

In Chapter 3, we present the result from a user study to determine if the presence of lyrics can help increase the agreement between Canadian and Chinese listeners about the mood of the song. We study the impact of the presence of lyrics on music mood perception for both Canadian and Chinese listeners by conducting a user study of Canadians not of Chinese origin, Chinese-Canadians, and Chinese people who have lived in Canada for fewer than three years. While our original hypotheses were largely connected to cultural components of mood perception, we also analyzed how stable mood assignments were when listeners could read the lyrics of recent popular English songs they were hearing versus when they only heard the songs. We find that the mood assigned to a song is dependent on whether people listen to it, read the lyrics or both together. We also showed the lyrics of some songs to participants without playing the recorded music. For example, people assign different moods to the same song in these three scenarios. People tend to assign a song to the mood cluster that includes “melancholy” more often when they read the lyrics without listening to it, and having access to the lyrics does not help reduce the difference in music mood perception between Canadian and Chinese listeners significantly. Our results cause us to question the idea that songs have inherent mood. Rather, we suggest that the mood depends on both cultural and experiential context.

1.2 Lyrics vs Poetry

In Chapter 4, we show that we can predict the genre of a document based on the adjective choices made by authors. We hypothesize that different genres of writing use different adjectives for the same concept. We test our hypothesis on lyrics, articles and poetry. We use the English Wikipedia and over 13,000 news articles from four leading newspapers for the article data set. Our lyrics data set consists of lyrics of more than 10,000 songs by 56 popular English singers, and our poetry dataset is made up of more than 20,000 poems from 60 famous poets. We find the probability distribution of synonymous adjectives in all the three different categories and use it to predict if a document is an article, lyrics or poetry given its set of adjectives. We achieve an accuracy level of 67% for lyrics, 80% for articles and 57% for poetry. Using this approach we show that adjectives more likely to be used in lyrics are more rhymable than those more likely to be used in poetry, but they do not differ significantly in their semantic orientations, which was found using SentiWordNet [30]. Furthermore we show that our algorithm is successfully able to detect “poetic” lyricists like

Bob Dylan, who have published books of poetry, from non-poetic ones like Bryan Adams, as the lyrics of more “poetic” lyricists are more often misclassified as poetry.

1.3 Hit Song Detection

In Chapter 5, we introduce our hit detection model. The music information retrieval task of predicting hits is largely unsolved [68]. Previous efforts to predict whether a song will be a hit have focused on audio features of the sound recording. We instead focus on the lyrics, which are an opportunity for songwriters to show off their artisanship, and which can be more easily analyzed using computer algorithms. Using 31 rhyme, syllable and meter features, we create Bayesian network and support vector machine filters that are surprisingly effective at separating hits from flops. We define hits as songs that made it to the Billboard Year-End Hot 100 singles chart between the years 2008 and 2013. Flops are harder to define: they are non-hit songs that had a chance of being hits, for example because of having had enough airplay to appear on a weekly chart, or by having been released by a singer with many hits. Since it is difficult to agree on the definition of flops, we analyze several variant definitions. Our largest data set consists of 492 hits and 6323 flops. Using cross validation, a weighted support vector machine gives us recall and precision values of 0.492 and 0.243 respectively for the hits on our largest data set, which is much stronger than would be expected by random chance. Adding fourteen audio features gives a slight improvement, but the lyrics features are significantly much more useful than audio features in separating hits and flops. We argue that complex rhyme and meter is a detectable property of lyrics that indicates quality songmaking, and that it is this property that allows our filter to predict hit songs successfully.

1.4 How To Get Shazamed?

In Chapter 6, we present a way for upcoming artists to rise above “sonic wallpaper” and become successful through the route of highly-shazamed lyrically complex songs. The music recognition service Shazam has been used to identify more than 15 billion songs, with over 500 million users. People use Shazam to identify songs they do not know about in settings where they cannot see the artist and title information. Songs with over 5 million shazams¹ are placed in the Shazam Hall of Fame [67]. We seek to identify what makes a

¹In this manuscript, we use “shazam” as both noun and verb, consistent with popular usage, and we write it with lower case.

song catchy enough for users to shazam it, focusing on complexity in the rhyme and meter pattern of songs. In particular we, seek to separate songs in the Shazam Hall of Fame from hit songs (not found in the Hall of Fame). Songs in the Shazam Hall of Fame are lyrically more complicated and in particular have more complex rhymes than hits. Using linear regression to predict the number of shazams, we show a model using the lyrical complexity as a feature better predicts the number of shazams. Additionally, we note that many of the songs in the Shazam Hall of Fame are by relatively unknown artists who use an early Shazam success to create visibility, and we conclude that, one way for an artist to break out is to write catchy complex lyrics with complicated rhymes. Songs by upcoming artists usually have the chorus as the lyrically least complex part of the song while songs by established artists usually have chorus as the lyrically most complex part of the song, and we conjecture that this may relate to how the song-writing process changes as a musician's career progresses.

1.5 Summary

In this thesis, we show the usefulness of lyrics in solving four important Music Information Retrieval tasks. For our first result, we show that the presence of lyrics has a variety of significant effects on how people perceive songs. We then proceed to show that we can predict the genre of a document based on the adjective choices made by the authors. For our next result, we show that rhyme and meter features are useful in separating hits and flops and outperform the currently popularly used audio features. Using the same features we can also detect songs which are likely to be shazamed heavily. We argue that complex rhyme and meter is a detectable property of lyrics that indicates quality songmaking and artisanship and allows artists to become successful.

The material in this thesis is based on our previous works, both published and under review. Chapter 3 and 4 is based on our papers which we presented at ISMIR 2014 ([78, 76]). The hit detection model in Chapter 5 is based on the late breaking demo we presented at ISMIR 2014 [77] and a paper we presented in CMMR 2015 [79].

Chapter 2

Background and Related Work

Lyrics have widely been ignored in Music Information Retrieval research. The main focus has been on the use of low level audio features to solve MIR problems. Lee et al. [54] analyze the topics of International Society for Music Information Retrieval, the top MIR conference, papers from 2000 to 2008. Analyzing the most commonly used title and abstract terms, they conclude that the focus of research has mainly been on audio. Lyrics does not make it to the top-10 ranked title terms for any of the years between 2000 and 2008. Downie et al. [27] analyze the first 10 years, 1999 to 2009, of International Society for Music Information Retrieval conference. They conclude that there was a heavy emphasis on music in symbolic form over audio during ISMIR's early years but audio is now the main focus of MIR research.

2.1 Lyrics and Neuroscience

Lyrics, though largely ignored [54], are an integral part of a listeners musical experience and can be useful in solving in important MIR tasks. Lyrics contain much of the emotional content of a typical popular song. Anderson et al. [2], examined effects of songs with violent lyrics on aggressive thoughts and hostile feelings. They demonstrated that college students who heard a violent song felt more hostile than those who heard a similar but nonviolent song. These effects replicated across songs and song types (e.g., rock, humorous, nonhumorous). Furthermore, behavioral and neuropsychological research has shown that individuals process lyrics and tune separately while listening to a song. Besson et al. [9], study whether people listening to a song treat the linguistic and musical components

separately or integrate them within a single percept. They find that harmonic processing is not affected by the semantics of the sentence even when presented in stimuli in which the lyrics and the tunes are strongly intertwined. They conclude that lyrics and tunes in vocal music may be integrated in memory, but they are processed independently on-line when the semantic and harmonic aspects are considered.

Stratton et al. [84] conducted experiments on college students to examine the relative impact of lyrics versus music on mood. They find that sad lyrics along with music increased depression and decreased positive affect, even for songs performed in an upbeat style. Furthermore, melodies paired with sad lyrics were rated as less pleasant when students heard the melody by itself. They conclude that lyrics appear to have greater power to direct mood change than music alone. Lennings and Warburton [55] ran an experiment where 194 participants heard music either with or without lyrics, and with or without a violent music video, and were then given the chance to aggress. They find that the strongest effect was elicited by exposure to violent lyrics, regardless of whether violent imagery accompanied the music, and regardless of various person-based characteristics.

Guéguen [34] et al. study the effect of romantic lyrics on 18 to 20 year old single female participants. They were made to hear romantic lyrics or neutral ones while waiting for the experiment to start. Five minutes later, the participant interacted with a young male confederate in a marketing survey. During a break, the male confederate asked the participant for her phone number. It was found that women previously exposed to romantic lyrics complied with the request more readily than women exposed to the neutral ones.

Lyrics contains much of the typical emotional content of a song and has a greater power to direct mood change than music alone. Neuropsychological research has shown that individuals process lyrics and tune separately. Hence, lyrics form a very vital component of any song and can be useful in solving challenging MIR tasks.

2.2 Lyrics and Music Information Retrieval

Past research has shown textual features to be better than audio features at solving certain MIR tasks. Hu and Downie [46], combine audio and text features for multi modal mood classification. Out of the 18 categories, textual features significantly outperformed the audio features in seven categories, while the audio features outperforms all textual features in only one category. Dhanaraj and Logan [24], use both text and audio features individually and together to predict hit songs. They learn the most prominent sounds and topics of each song (using textual analysis), and conclude that the text features are slightly

more useful than the audio features; combining both of them together does not produce significant improvements. Combining textual and audio features can be useful at times. Hu and Downie [45], evaluated the usefulness of textual features in music mood classification. They conclude that systems using both the audio and textual features outperformed systems just using the audio features in mood classification.

Lyrics features like rhyme and meter, sets lyrics and poetry apart from prose and can be useful in solving difficult MIR problems. Hirjee and Brown [40], came up with a method of scoring potential rhymes using a probabilistic model based on phoneme frequencies in rap lyrics. They used this scoring to automatically identify internal and line-final rhymes in song lyrics. They conclude that their probabilistic method is superior at detecting rhymes than the rule based methods. Hirjee and Brown [41], developed a probabilistic model of misheard lyrics trained on actual misheard lyrics, and develop a phoneme similarity scoring matrix based on this model. They conclude that the probabilistic method is superior to other methods at finding the correct lyrics.

Smith et al. [80] used tf-idf weighting to find typical phrases and rhyme pairs in song lyrics. They develop an application that estimates how clichéd a song is and conclude that the typical number-one hits, on average, are more clichéd. They believe that song popularity and writing quality are not necessarily connected.

Previous work on poetry has focused on poetry translation, and automatic poetry generation. Genzel et al. [33] develop a system for the machine translation of poetry. They show that a machine translation system can be constrained to search for translations obeying a particular length, meter and rhyming constraint. However, the impact on translation quality is profound and the system is too slow. Jiang and Zhou [51] generate Chinese couplets using statistical machine translation. The system takes the first sentence as the input and generates the N-best list of proposed second sentences. They filter the candidates that violate the linguistic constraints and rank the remaining candidates using a support vector machine.

Lyrics, compared to the audio, is currently largely ignored in the MIR research, and in this thesis we show that it is possible to solve some traditionally difficult MIR problems like hit song detection using lyric features. We believe that lyrics analysis has the potential to solve other challenging MIR problems like: playlist generation, song segmentation, genre and mood detection. High-level lyric features can be combined with low level audio features to improve the existing benchmarks of important MIR problems.

Chapter 3

On Cultural, Textual And Experiential Aspects Of Music Mood

3.1 Introduction

Music mood detection has been identified as an important Music Information Retrieval (MIR) task. It is based on the belief that every song has an inherent mood. It is an important feature of music recommendation systems as mood is an important criteria based on which people search for songs [94]. Though most automatic mood classification systems are solely based on the audio content of the song, some systems have used lyrics or have combined audio and lyrics features (e.g., [46, 36, 49] and [45, 53]) Previous studies have shown that combining these features improves classification accuracy (e.g., [45, 53] and [95]) but as mentioned by Downie et al. in [46], there is no consensus on whether audio or lyrical features are more useful.

Implicit in mood identification is the belief that songs have inherent mood, but in practice this assignment is unstable. Recent work has focused on associating songs with more than one mood label, where similar mood tags are generally grouped together into the same label (e.g.,[88]), but this still tends to be in a stable listening environment.

Our focus is instead on the cultural and experiential context in which people interact with a work of music. People's cultural origin may affect their response to a work of art, as may their previous exposure to a song, their perception of its genre, or the role that a song or similar songs has had in their life experiences. We focus on people's cultural origin, and on how they interact with songs (for example, seeing the lyrics sheet or not). Listening to

songs while reading lyrics is a common activity: for example, there are lyrics videos (which only show lyrics text) on YouTube with hundreds of millions of views (e.g. Green Day’s “Boulevard of Broken Dreams”) [66], and CD liner notes often include the text of lyrics [12]. Our core hypothesis is that there is enough plasticity in assigning moods to songs, based on context, to argue that many songs have no inherent mood.

Past studies have shown that there exist differences in music mood perception among Chinese and American listeners (e.g., [48]). We surmised that some of this difference in mood perception is due to weak English language skills of Chinese listeners: perhaps such listeners are unable to grasp the wording in the audio. We expected that they might more consistently match the assignments of native English-speaking Canadians when shown the lyrics to songs they are hearing than in their absence. We addressed the cultural hypothesis by exploring Canadians of Chinese origin, most of whom speak English natively but have been raised in households that are at least somewhat culturally Chinese. If such Chinese-Canadians match Canadians not of Chinese origin in their assignments of moods to songs, this might at least somewhat argue against the supposition that being Chinese culturally had an effect on mood assignment, and would support our belief that linguistic skills account for at least some of the differences. Our campus has many Chinese and Chinese-Canadians, which also facilitated our decision to focus on these communities.

In this study we use the same five mood clusters as are used in the MIREX audio mood classification task and ask the survey participants to assign a song to only one mood cluster. A multimodal mood classification could be a possible extension to our work here. Some in MIR work [52] had used Russell’s valence-arousal model, where the mood is determined by the valence and arousal scores of the song; we use the simpler 5-group classification here.

In practice, our hypotheses about language expertise were not upheld by our experimental data. Rather, our data support the claim that both cultural background and experiential context have significant impact on the mood assigned by listeners to songs, and this effect makes us question the meaningfulness of mood as a category in MIR.

This work was published at ISMIR 2014 [78], and this chapter quotes extensively from that publication.

3.2 Related Work

3.2.1 What is Mood Detection?

Mood classification is a classic task in MIR, and is based on the belief that every song has an inherent mood. The aim of the music mood detection algorithms is to detect the inherent mood of the song rather than the mood induced by the song, which is more subjective. The Music Information Retrieval Evaluation eXchange (MIREX) [26] is a community-based formal evaluation framework to create the necessary infrastructure for the scientific evaluation of the many different techniques being employed by researchers interested in the domains of Music Information Retrieval. Audio music mood classification happens to be an important MIREX challenge.

3.2.2 Methods for Detecting Mood

Audio analysis has been the primary focus of different mood detection projects. Lyrics have largely been neglected and at times have been used along with the audio features. Lu et al. [56] and Trohidis et al. [88] come up with an automatic mood classification system solely based on audio. Several projects like Downie et al. [46], Xiong et al. [36] and Chen et al. [49], have used lyrics as part of the mood prediction task. Downie et al. [46] show that features derived from lyrics outperform audio features in seven out of the eight categories. Downie et al. [45], Laurier et al. [53] and Yang et al. [95] show that systems which combine audio and lyrics features outperform systems using only audio or only lyrics features. Downie et al. [45] show that using a combination of lyrics and audio features reduces the need of training data required to achieve the same or better accuracy levels than only-audio or only-lyrics systems.

3.2.3 Mood Tags

Downie et al. [46], Laurier et al. [53] and Lee et al. [48] use 18 mood tags derived from social tags and use multimodal mood classification system. Trohidis et al. [88] use multi modal mood classification into six mood clusters. Kosta et al. [52] use Russell’s valence-arousal model which has 28 emotion denoting adjectives in a two dimensional space. Downie et al. [44] use the All Music Guide datasets to come up with 29 mood tags and cluster it into five groups. These five mood clusters, shown in Table 3.1, are used in the MIREX audio

music mood classification task. We use these clusters, where each song is assigned a single mood cluster.

3.2.4 Music Mood Perception between different Cultures

Lee et al. [48] study the difference in music mood perception between Chinese and American listeners on a set of 30 songs and conclude that mood judgment differs between Chinese and American participants and that people belonging to the same culture tend to agree more on music mood judgment. That study primarily used the common Beatles data set, which may have been unfamiliar to all audiences, given its age. Their study collected mood judgments solely based on the audio; we also ask participants to assign mood to a song based on its lyrics or by presenting both audio and lyrics together. To our knowledge, no work has been done on the mood of a song when both audio and lyrics of the song is made available to the participants, which as we have noted is a common experience. Kosta et al. [52] study if Greeks and non-Greeks agree on arousal and valence rating for Greek music. They conclude that there is a greater degree of agreement among Greeks compared to non-Greeks possibly because of acculturation to the songs.

Cluster 1	passionate, rousing, confident, boisterous, rowdy
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster 3	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 3.1: The mood clusters used in the study.

3.3 Method

3.3.1 Data Set

We selected fifty very popular English-language songs of the 2000s, with songs from all popular genres, and with an equal number of male and female singers. We verified that the selected songs were international hits by going to the songs’ Wikipedia pages and analyzing the peak position reached in various geographies. The list of songs, their artists and the year of release are in Table 3.3.

Song	Artist	Year of Release
Umbrella	Rihanna	2007
American Idiot	Green Day	2004
Beautiful Day	U2	2000
Oops I did it Again	Britney Spears	2000
Party in the USA	Miley Cyrus	2010
Pocketful of Sunshine	Natasha Bedingfield	2010
Hips Don't Lie	Shakira	2005
Hero	Enrique Iglesias	2001
In the End	Linkin Park	2000
It's not Over	Daughtry	2006
You're Beautiful	James Blunt	2004
Maria Maria	Santana	2000
Human	The Killers	2008
We Belong Together	Mariah Carey	2005
Bad Day	Daniel Powter	2005
London Bridge	Fergie	2006
Bleeding Love	Leona Lewis	2007
Viva La Vida	Coldplay	2008
Disturbia	Rihanna	2007
Womanizer	Britney Spears	2008
When I'm Gone	Eminem	2005
Whenever Wherever	Shakira	2001
Yellow	Coldplay	2000
Bubbly	Colbie Caillat	2007
Complicated	Avril Lavigne	2002
Apologize	One Republic	2007
You Sang to me	Marc Anthony	2000
It's My Life	Bon Jovi	2000
Boulevard of Broken Dreams	Green Day	2004
Poker Face	Lady Gaga	2008
Feel	Robbie Williams	2002
Rolling in the Deep	Adele	2011
Irreplaceable	Beyonce	2006
Don't Phunk with my Heart	The Black Eyed Peas	2005
Stars are Blind	Paris Hilton	2006

Sexy and I Know it	LMFAO	2011
We Like to Party	Vangaboys	1998
Numb	Linkin Park	2003
Smack That	Akon	2006
Love Story	Taylor Swift	2008
No Matter what	Boyzone	1998
The Ketchup Song	Las Ketchup	2002
Wake Up	Hillary Duff	2007
Single Ladies (Put a Ring on It)	Beyonce	2008

Table 3.3: The list of names, artists and the year of release of songs which were used in the study.

We focus on English-language popular music in our study, because it is the closest to “universally” popular music currently extant, due to the strength of the music industry in English-speaking countries. Our data set includes music from the US, Canada, the UK, and Ireland.

3.3.2 Participants

The presence of a large Chinese and Canadian population at our university, along with obvious cultural differences between the two communities, convinced us to use them for the study. We also include Canadians of Chinese origin; we are unaware of any previous MIR work that has considered such a group. We note that the Chinese-Canadian group is diverse: while some speak Chinese languages, others have comparatively little exposure to Chinese language or culture [21]. We recruited 100 participants, mostly university students, from three groups. Our Chinese group consisted of 33 Chinese, living in Canada for less than 3 years. Our second group consisted of 33 Canadians, not of Chinese origin, born and brought up in Canada, with English as their mother tongue. Our final group was made up of 34 Canadians of Chinese origin, born and brought up in Canada.

3.3.3 Survey

Each participant was asked to assign a mood cluster to each song in a set of 10 songs. For the first three songs they saw only the lyrics; for the next three songs they only heard

the first 90 seconds of the audio; and for the last four songs they had access to both the lyrics and the first 90 seconds of the audio simultaneously. They assigned each song to one of the five mood clusters shown in Table 3.1. We collected 1000 music mood responses for 50 songs, 300 each based solely either on audio or lyrics and 400 based on both audio and lyrics together. We note that due to their high popularity, some songs shown only via lyrics may have been known to some participants. We did not ask participants if this was the case.

3.4 Results

We hypothesized that the difference in music mood perception between American and Chinese listeners demonstrated by Hu and Lee [48] is because of the weak spoken English language skills of Chinese students, and that this might give them some difficulty in understand the wording of songs; this is why we allowed our participants to see the lyrics for seven out of ten songs. Hence, we had hypothesized before our study that Chinese-born Chinese will more consistently match Canadians when they are shown the lyrics to songs, and Chinese-born Chinese listeners will have less consistency in their assignment of moods to songs than do Canadian-born non-Chinese when given only the recording of a song. Furthermore, we believed that just reading the lyrics will be more helpful in matching Canadians than just hearing the music for Chinese-born Canadians. We believed that Canadian-born Chinese participants will be indistinguishable from Canadian-born non-Chinese participants since they have similar English Language skills. We also believed mood to be dependent on experiential context and had hypothesized that people often assign different mood to the same song depending on whether they read the lyrics, or listen the audio or both simultaneously. Finally, since music mood is heavily dependent on experiential and cultural context we believed that a song does not have an inherent mood: its “mood” depends on the way it is perceived by the listener, which is often listener-dependent.

3.4.1 Lyrics and music mood perception between cultures

Hu and Lee [48] had shown that there exists difference in music mood perception between American and Chinese listeners. Hence, we started this study with the hypothesis that difference in music mood perception between Chinese and Canadian cultures is partly caused by English language skills, since the spoken English-language skill of Chinese people are weaker than that of Canadians, and that if participants are asked to assign mood to

a song based on its lyrics, we will see much more similarity in judgment between the two different groups.

We used the Kullback-Leibler distance, which is a non-symmetric measure of the difference between two probability distributions, between the distribution of responses from one group and the distribution of responses from that group and another group to identify how similar the two groups' assignments of moods to songs were, and we used a permutation test to identify how significantly similar or different the two groups were. We ran the permutation test 1000 times and checked for statistical significance at a p value of 1%. In Table 3.4, we show the number of songs for which different population groups are surprisingly similar. What we find is that the three groups actually somewhat agree in uncertainty of assigning mood to songs when they are presented only with the recording: if one song has uncertain mood assignment for Canadian listeners, our Chinese listeners also typically did not consistently assign a single mood to the same song.

Our original hypothesis was that adding presented lyrics to the experience would make Chinese listeners agree more with the Canadian listeners, due to reduced uncertainty in what they were hearing. In actuality, this did not happen at all: in fact, presence of both audio and lyrics resulted in both communities having both more uncertainty and disagreeing about the possible moods to assign to a song.

This confusion in assigning a mood might be because a lot of hit songs (Green Day's "Boulevard of Broken Dreams", Coldplay's "Viva La Vida", James Blunt's "You're Beautiful", *etc.*) use depressing words with very upbeat tunes. It could also be that by presenting both lyrics and audio changes the way a song is perceived by the participants and leads to a completely new experience. (We note parenthetically that this argues against using lyrics-only features in computer prediction of song mood, as listeners do seem to, themselves, respond incompletely with only the words.)

The number of songs with substantial agreement between Chinese and Canadian, not of Chinese origin, participants remains almost the same with lyrics only and audio only, but falls drastically when both are presented together. (Note again: in this experiment, we are seeing how much the distribution of assignments differs for the two communities.) This contradicts our hypothesis that the difference in music mood perception between Chinese and Canadians is because of their difference in English abilities. It could of course be the case that many Chinese participants did not understand the meaning of some of the lyrics.

We had hypothesized that Canadians of Chinese and non-Chinese origin would have very similar mood judgments because of similar English language skills but they do tend to disagree a lot on music mood. The mood judgment agreement between Chinese and Canadians of Chinese and non-Chinese origin seem to be similar (permutation test at

$p > 0.99$) and we conclude that we can make no useful claims about the Chinese-Canadian participants in our sample.

On the whole we conclude that the presence of lyrics does not significantly increase the music mood agreement between Chinese and Canadian participants: in fact, being able to read lyrics while listening to a recording seems to significantly decrease the music mood agreement between the groups.

		lyrics	audio	audio+lyrics
Chinese	Canadians	25	22	14
Chinese	Canadian-Chinese	36	31	27
Chinese	non-Chinese Canadians	31	32	23
non-Chinese Canadians	Canadian-Chinese	36	29	31

Table 3.4: The number of statistically-significantly similar responses between the different cultures for the three different ways they interact with the songs. Canadians refer to Canadians of both Chinese and non-Chinese origin.

3.4.2 Stability across the three kinds of experiences

We analyze the response from participants when they are made to listen to the lyrics, hear the audio or both simultaneously across all the three groups. We calculate Shannon entropy of this mood assignment for each of the 50 songs for the three ways we presented a song to the participants: some songs have much more uncertainty in how the participants assign mood cluster to them. We then see if this entropy is correlated across the three kinds of experience, using Spearman’s rank correlation coefficient of this entropy value between the groups. A rank correlation of 1.0 would mean that the song with the most entropy in its mood assignment in one experience category, that is reading the lyrics or listening to the audio or both together, is also the most entropic in the other, and so on. The Spearman’s correlation coefficients can be found in Table 3.5.

	Spearman’s rank correlation coefficient
only lyrics versus only audio	0.0504
only lyrics versus audio plus lyrics	0.1093
only audio versus audio plus lyrics	0.0771

Table 3.5: Spearman’s rank correlation coefficient between the groups. The groups “only lyrics” and “only audio” identify when participants had access to only lyrics and audio respectively while “audio+lyrics refers to when they had access to both simultaneously.

The low value of the correlation analysis suggests that there is almost no relationship between “certainty” in music mood across the three different kinds of experiences: for songs like “Wake Up” by Hillary Duff and “Maria Maria” by Santana, listeners who only heard the song were consistent in their opinion that the song was from the second cluster, “cheerful”, while listeners who heard the song and read the lyrics were far more uncertain as to which class to assign the song to.

3.4.3 “Melancholy” lyrics

For each song, we identify the mood cluster to which it was most often assigned, and show these in Table 3.6.

Mood Clusters	Example Word	only lyrics	only audio	audio plus lyrics
Cluster 1	Passionate	8	9	13
Cluster 2	cheerful	5	15	11
Cluster 3	poignant	28	14	18
Cluster 4	humorous	4	6	3
Cluster 5	aggressive	5	6	5

Table 3.6: The most commonly assigned mood clusters for each experimental context. Most songs are assigned to the third mood cluster when participants are shown only the lyrics.

Songs experienced only with the lyrics are most often assigned to the third mood cluster, which includes the mood tags similar to “melancholy”. In the presence of audio or both audio and lyrics there is a sharp decline in the number of songs assigned to that cluster; this may be a consequence of “melancholy” lyrics being attached to surprisingly cheery tunes, which cause listeners to assign them to the first two clusters. The number of songs assigned to the fourth and fifth cluster remains more similar across all experiential contexts. Even between the two contexts where the listener does hear the recording of the song, there is a good deal of inconsistency in assignment of mood to songs: for 27 songs, the most commonly identified mood is different between the “only audio” and “audio+lyrics” data.

3.4.4 Genre and Mood

We explored the different genres in our test set, to see if our different cultural groups might respond in predictable ways when assigning moods to songs.

3.4.4.1 Rock songs

Things that might be considered loud to Chinese listeners could be perceived as normal to Canadian listeners due to their cultural differences [48]. Thus, we examined how responses differed across these two groups for rock songs, of which we had twelve in our data set. We calculate the Shannon entropy of the response of the participants and present the result in Table 3.7. We see that for many rock songs, there is high divergence in the mood assigned to the song by our listeners from these diverse cultures. For seven of the twelve rock songs, the most diversity of opinion is found when listeners both read lyrics and hear the audio, while for three songs, all three participants who only read the lyrics agreed exactly on the song mood (zero entropy).

We see that for three of twelve cases all the participants tend to agree on the mood for the song when they are given access to the lyrics. The data for lyrics only have lower entropy than audio for five of twelve cases and all five of these songs are “rebellious” in style. For the five cases where the audio-only set has lower entropy than lyrics-only, the song has a more optimistic feel to it. This is consistent with our finding in the last section about melancholy song lyrics.

For example, the lyrics of “Boulevard of Broken Dreams”, an extremely popular Green Day song, evoke isolation and sadness, consistent with the third mood cluster. On the other hand the song’s music is upbeat which may give the increased confusion when the participant has access to both the audio and lyrics for the song.

3.4.4.2 Hip-Hop/ Rap

Lee et al. [48] show that mood agreement among Chinese and American listeners is least for dance songs. They have four instrumental dance songs in their data set and see an agreement ratio of 0.22 between American and Chinese listeners. The agreement ratio between two listeners will be 1 if they agree on the mood of all the songs and 0 if disagree about the mood of every song. Our test set included five rap songs, and since this genre is often used at dance parties, we analyzed user response for this genre. Again, we show the entropy of mood assignment for the three different experiential contexts in Table 3.8.

What is again striking is that seeing the lyrics (which in the case of rap music is the primary creative element of the song) creates more uncertainty among listeners as to the mood of the song, while just hearing the audio recording tends to yield more consistency. Perhaps this is because the catchy tunes of most rap music pushes listeners to make a spot judgment as to mood, while being reminded of lyrics pushes them to evaluate more complexity.

In general we see that there is high entropy in mood assignment for these songs, and so we confirm the previous claim that mood assignment is less certain for “danceable” songs.

Song	only lyrics	only audio	audio+lyrics
“Complicated”	1.0	0.918	1.148
“American Idiot”	1.792	1.459	1.792
“Apologize”	1.0	1.25	1.0
“Boulevard of Broken Dreams”	0.0	1.792	2.155
“Bad Day”	1.792	1.459	1.061
“In the End”	0.65	1.459	1.061
“Viva La Vida”	0.0	1.5849	1.75
“It’s My life”	0.0	0.65	1.298
“Yellow”	1.792	0.65	1.351
“Feel”	0.918	0.650	1.148
“Beautiful Day”	1.584	1.459	1.836
“Numb”	1.25	1.918	0.591

Table 3.7: Entropy values for the mood assignment of rock songs, for the three different categories of interactions with a song. An entropy value of 0 indicates that everyone agreed on the mood of the song.

Song	only lyrics	only audio	audio+lyrics
“London Bridge”	1.459	0.918	1.405
“Dont Phunk With My Heart”	1.459	1.251	1.905
“I Wanna Love You”	0.918	1.459	1.905
“Smack That”	1.918	1.792	1.905
“When I’m Gone”	1.251	0.918	1.448

Table 3.8: Entropy values for hip-hop/ rap songs for the three different categories for each of the three ways people interact with a song. An entropy value of 0 indicates that everyone agreed on the mood of the song.

3.5 Conclusions: Does Music Mood Exist?

For music mood classification to be a well-defined task, the implicit belief is that songs have “inherent mood(s),” that are detectable by audio features. Our hypothesis is that

many songs have no inherent mood, but that the perceived mood of a song depends on cultural and experiential factors. The data from our study supports our hypothesis.

We have earlier shown that the mood judgment of a song depends on whether it is heard to or its lyrics is read or both together, and that all three contexts produce mood assignments that are strikingly independent.

We have shown that participants are more likely to assign a song to the “melancholy” mood cluster when only reading its lyrics, and we have shown genre-specific cultural and experiential contexts that affect how mood appears to be perceived. Together, these findings suggest that that the concept of music mood is fraught with uncertainty.

The MIREX audio mood classification task has had a maximum classification accuracy of less than 70% [47], with no significant recent improvements. Perhaps, this suggests that the field is stuck at a plateau, and we need to redefine “music mood” and change our approach to the music mood classification problem. Music mood is highly affected by external factors like the way a listener interacts with the song, the genre of the song, the mood and personality of the listener, and future systems should take these factors into account.

Chapter 4

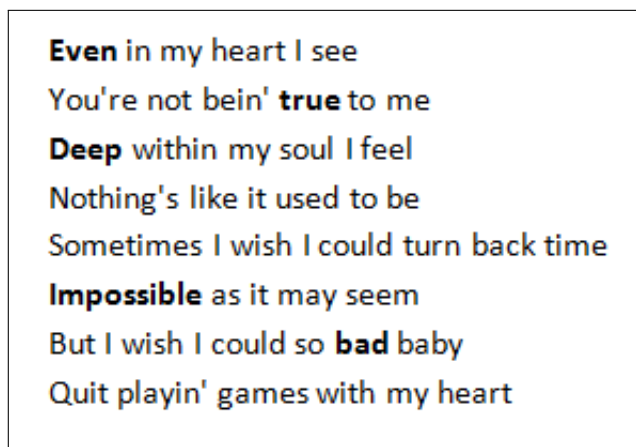
Are Poetry and Lyrics All That Different?

4.1 Introduction

The choice of a particular word, from a set of words that can all be used appropriately, depends on the context we use it in, and on the artistic decision of the authors. We believe that for a given concept, the words that are more likely to be used in lyrics will be different from the ones which are more likely to be used in articles or poems, because lyricists typically have different objectives [91]. Poetry typically attracts a more educated and sensitive audience while lyrics are written for the masses and hence the tendency to use simpler words in music lyrics and comparatively more sophisticated words in poetry [70]. We test our hypothesis by examining adjective usage in these categories of documents. We focus on adjectives, as a majority have synonyms that can be used depending on context. To our surprise, we find adjective usage is sufficient to separate lyrics from poetry quite effectively.

4.1.1 Definitions and Synonyms

Finding the synonyms of a word is still an open problem [93]. We used three different sources to obtain synonyms for a word: the WordNet [59], Wikipedia [86] and an online thesaurus [87]. We prune synonyms, obtained from the three sources, which fall below an experimentally determined threshold for the semantic distance (calculated using the



Even in my heart I see
You're not bein' **true** to me
Deep within my soul I feel
Nothing's like it used to be
Sometimes I wish I could turn **back** time
Impossible as it may seem
But I wish I could so **bad** baby
Quit playin' games with my heart

Figure 4.1: The bold-faced words are the adjectives our algorithm takes into account while classifying a document, which in this case is a snippet of lyrics by the Backstreet Boys.

method described by Pirro et al. [69]) between the synonyms and the word. The list of relevant synonyms obtained after pruning was used to obtain the probability distribution over words for each class of document, for common adjectival concepts. Semantic distance is a metric defined over a set of terms, where the idea of distance between them is based on the likeness of their meaning or semantic content. We use the method described by Pirro et al. [69], where they consider a broader range of relations (e.g., part-of) along with assessing how two objects are alike.

Lyrics and poetry started as very similar concepts. In the early nineteenth century, lyric was one of three broad categories of poetry in classical antiquity, along with drama and epic [16]. Lyric poetry was a form of poetry which expresses personal emotions or feelings, typically spoken in the first person usually with a musical accompaniment known as a lyre [16]. Over the course of time both have evolved into two slightly different elements. A key requirement of our study is that there exists a difference, albeit a hazy one, between poetry and lyrics. Poetry attracts a more educated and sensitive audience while lyrics are written for the masses [70]. Poetry, unlike lyrics, is often structurally more constrained, adhering to a particular meter and style [70]. Lyrics are often written keeping the music in mind while poetry is written against a silent background. Lyrics, unlike poetry, often repeat lines and segments, causing us to believe that lyricists tend to pick more rhymable adjectives. Furthermore, a good song made by its hooks, and there are several different kinds of hook. One of them is catchy rhyme, which can contribute to the memorizability of songs [89]. Of course, some poetic forms also repeat lines, such as the villanelle.

We use a bag of words model for the adjectives, where we do not care about their relative positions in the text, but only their frequencies. Finding synonyms of a given word is a vital step in our approach and since it is still considered a difficult task, improvement in synonyms finding approaches will lead to an improvement in our classification accuracy.

Our classification algorithm has a linear run time as it scans through the document once to detect the adjectives and calculate the probability of the document being a poetry, lyric or an article. The document class with the highest probability is chosen as the outcome. We attain an overall accuracy of 68%. Lyricists with a relatively high percentage of lyrics misclassified as poetry tend to be recognized for their poetic style, such as Bob Dylan, who has published books of poetry, and Annie Lennox.

This work was published at ISMIR 2014 [76], and this chapter quotes extensively from that publication.

4.2 Related Work

We do not know of any work on the classification of documents based on the adjective usage into genre, nor are we aware of any computational work which discerns poetic from non-poetic lyricists. Previous works have used adjective choice for various purposes like sentiment analysis [20]. Work on poetry has focused on poetry translation, automatic poetry generation, rather than focusing on the word choice of poetry.

Chesley et al. [20] classifies blog posts according to sentiment using verb classes and adjective polarity, achieving accuracy levels of 72.4% on objective posts, 84.2% for positive posts, and 80.3% for negative posts. Entwisle et al. [29] analyzes the free verbal productions of ninth-grade males and females and conclude that girls use more adjectives than boys but fail to reveal differential use of qualifiers by social class.

Smith et al. [80] use tf-idf weighting to find typical phrases and rhyme pairs in song lyrics and conclude that the typical number one hits, on average, are more clichéd. Nichols et al. [64] studies the relationship between lyrics and melody on a large symbolic database of popular music and conclude that songwriters tend to align salient (prominent) notes with salient (prominent) lyrics.

There is some existing work on automatic generation of synonyms. Zhou et al. [93] extracts synonyms using three sources - a monolingual dictionary, a bilingual corpus and a monolingual corpus, and use a weighted ensemble to combine the synonyms produced from the three sources. They get improved results when compared to the manually built

thesauri, WordNet [59] and Roget [74]. Christian et al. [14] describe an approach for using Wikipedia to automatically build a dictionary of named entities and their synonyms. They were able to extract a large amount of entities with a high precision, and the synonyms found were mostly relevant, but in some cases the number of synonyms was very high. Niemi et al. [65] add new synonyms to the existing synsets of the Finnish WordNet using Wikipedias links between the articles of the same topic in Finnish and English.

As to computational poetry, Jiang et al. [51] use statistical machine translation to generate Chinese couplets while Genzel et al. [33] use statistical machine translation to translate poetry keeping the rhyme and meter constraints. There is, a wide literature on generation of novel computational poetry, which we do not survey here.

4.3 Data Set

Artist	Poets
Bryan Adams	William Blake
Adele	E.E. Cummings
Akon	Edward FitzGerald
Beyonce	Robert Frost
Backstreet Boys	Donald Hall
Darius	Erica Jong
Green Day	John Keats
Celine Dion	Robert Lowell
Eminem	Walter De La Mare
Fergie	Adrienne Rich
Lady Gaga	Walter Scott
Enrique Iglesias	William Shakespeare
Rihanna	Percy Bysshe Shelley
Shakira	Lord Tennyson
U2	Alice Walker

Table 4.1: Fifteen of the singers and poets in our data set.

The training set consists a collection of of articles, lyrics and poetry and is used to calculate the probability distribution of adjectives in the three different types of documents. We use these probability distributions in our document classification algorithms, to identify

poetic from non-poetic lyricists and to determine adjectives more likely to be used in lyrics rather than poetry and vice versa.

4.3.1 Articles

We take the English Wikipedia and over 13,000 news articles from four major newspapers: The Chicago Sun-Times, Washington Post, Los Angeles Times, and The Hindu (which has editions from many Indian cities), as our article data set. Wikipedia, an enormous and freely available data set is edited by experts. Both of these are extremely rich sources of data on many topics. To remove the influence of the presence of articles about poems and lyrics in Wikipedia we ensured that the articles were not about poetry or music by not selecting articles belonging to the Entertainment or Music category to be in our data set.

4.3.2 Lyrics

We took more than 10,000 lyrics from 56 very popular English singers. The author and his supervisor both listen to English music and hence it was easy to come up with a list which included singers from many popular genres with diverse backgrounds. We focus on English-language popular music in our study, because it is the closest to “universally” popular music, due to the strength of the music industry in English-speaking countries. We do not know if our work would generalize to non-English Language songs. Our data set includes lyrics from American, Canadian, British and Irish lyricists. A list of fifteen of these singers is in Table 4.1.

4.3.3 Poetry

We took more than 20,000 English-language poems from 61 famous poets, like Robert Frost, William Blake and John Keats, over the last three hundred years. We selected the top poets from Poem Hunter [50]. A list of fifteen of these poets is in Table 4.1. We selected a wide time range for the poets, as many of the most famous English poets are from that time period. None of the poetry selected were translations from another language. Most of the poets in our dataset are poets from North America and Europe.

4.3.4 Test Data

Poetic Lyricists	Justification
Bob Dylan	Published poetry books [83]
Ed Sheeran	Writes poetic lyrics [90]
Ani Di Franco	A published poet [4]
Annie Lennox	Writes poetic lyrics [5]
Bill Callahan	Writes poetic lyrics [72]
Bruce Springsteen	Writes poetic lyrics [15]
Stephen Sondheim	Writes poetic lyrics [85]
Morrissey	Writes poetic lyrics [22]

Table 4.2: The list of poetic lyricists in our test set.

For the purpose of document classification we took 100 examples from each category, ensuring that they were not present in the training set. While collecting the test data we ensured the diversity, the lyrics and poets came from different genres and artists and the articles covered different topics and were selected from different newspapers. The lyricists in our test data consisted of artists like Lionel Richie, Kesha, Flo Rida, and INXS. The articles in our test set came from The New York Times, The Canadian Broadcasting Corporation and The Huffington Post. William Ernest Henley, James Joyce, Pablo Neruda and Thomas Hardy made up our poetry test set.

To determine poetic lyricists from non-poetic ones we took eight of each of the two types of lyricists, none of whom were present in our lyrics data sets. We ensured that the poetic lyricists we selected were indeed poetic by looking up popular news articles or ensuring that they were poet along with being lyricists. A few of our selected poetic lyricists like Bob Dyan and Ani DiFranco are published poets. Our list for poetic lyricists included Stephen Sondheim and Annie Lennox, while the non-poetic ones included Bryan Adams and Michael Jackson. The list of poetic lyricists in our test set is in Table 4.2.

4.4 Method

We start by finding the synonyms of all the adjectives in our training data set. We then proceed to calculate the probability distribution of adjectives in articles, lyrics and poetry. Using these probability distributions and synonyms generated in the previous steps, we

calculate the probability of a document in our test set being an article, lyrics or poetry and the predicted genre of the document is the highest probability choice.

4.4.1 Extracting Synonyms

We extract the synonyms for a term from three sources: WordNet, Wikipedia and an online thesaurus.

WordNet [59] is a large lexical database of English where words are grouped into sets of cognitive synonyms (synsets) together based on their meanings. WordNet interlinks not just word forms but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. The synonyms returned by WordNet need some pruning. For example, for “happy”, the WordNet returns “prosperous”, “halcyon”, “bright” and “golden” as synonyms, along with other plausible synonyms.

We use **Wikipedia** [86] redirects to discover terms that are mostly synonymous. It returns a large number of words, which might not be synonyms, so we need to prune the results. This method has been widely used for obtaining the synonyms of named entities (e.g. [14]), but we get decent results for adjectives too. Using the Wikipedia redirects to obtain the synonym of “happy”, we obtain “happyness”, “warm” and “fuzzy” and “felicitous” along with some other plausible synonyms.

We also used an **online thesaurus** [87] that lists words grouped together according to similarity of meaning. Though it gives very accurate synonyms, pruning is necessary to get better results. For “happy”, an online thesaurus returns “heaven-sent”, “tickled”, and “queer” along with the other plausible synonyms.

The semantic distance between each of the synonyms obtained from the three sources and the word was calculated. We prune synonyms which fall below a certain semantic similarity threshold, which was determined experimentally. Semantic distance is a metric defined over a set of terms, where the idea of distance between them is based on the likeness of their meaning or semantic content. To calculate the semantic similarity distance between words we use the method described by Pirro et al. [69], where they consider a broader range of relations (e.g., part-of) along with assessing how two objects are alike. Extracting synonyms for a given word is an open problem and with improvement in this area our algorithm will achieve better classification accuracy levels.

For example, for “happy” we obtain the following synonyms:

Via **WordNet**: elated, cheerful, happy, blessed, prosperous, golden, joyful, bright, laughing, riant, contented, glad, content, felicitous, halcyon, blissful, euphoric, joyous.

Via **Wikipedia redirects**: happiness, enjoyment, happy, gladness, lightheartedness, jolly, light-hearted, light-hearted, happyness, happier, warm, cheerfulness, happy, felicitous, felicitously, jocund, jocundity, jocundly, exulting, exulted, exults, exultantly, exultance, exultancy, exultingly, hapiness, jolliness, gaiety, happiest, happily.

Via **online thesaurus**: fluky, fortuitous, heaven-sent, lucky, providential, blissful, chuffed, delighted, gratified, joyful, joyous, pleased, satisfied, thankful, tickled, contented, gratified, happy, pleased, satisfied, fortunate, happy, applicable, appropriate, apt, becoming, befitting, felicitous, fitted, fitting, good, happy, meet, pretty, proper, right, suitable, happy, obsessed, queer.

After pruning: blessed, glad, felicitous, suitable, appropriate, blissful, good, meet, fitting, riant, joyful, laughing, lucky, euphoric, joyous, happy, fortuitous, pleased, cheerful, providential, elated.

4.4.2 Probability Distribution

We believe that the choice of an adjective to express a given concept depends on the genre of writing: adjectives used in lyrics will be different from ones used in poems or in articles. Poetry is typically written for a more sophisticated and educated audience while lyrics are written for the masses [70]. Hence, it is plausible to assume that lyrics will contain simpler words than poetry for the same concept. We calculate the probability of a specific adjective for each of the three document types.

First, WordNet is used to identify the adjectives in our training sets. For each adjective, we compute its frequency in the training set for all the three document classes and its synonyms. We then compute the frequency of all its synonyms in all the document classes. The ratio of the frequency of the adjective to the frequency of all its synonyms in all the document classes is calculated, which is the frequency with which that adjective represents its synonym group in that class of writing. This ratio is the probability of a document belonging to a particular document class given the occurrence of the adjective.

4.4.3 Document classification algorithm

We use a simple linear time algorithm which takes as input the probability distributions for adjectives, calculated above, and the document(s) to be classified, calculates the score

of the document being an article, lyrics or poetry, and labels it with the class with the highest score. The algorithm takes a single pass along the whole document and identifies adjectives using WordNet.

For each adjective in the document we check its presence in our training set for the particular document class. If found, we add the probability of the word in the given document class to the score, with a special penalty of -1 for adjectives never found in the training set for the document class and a special bonus of +1 for words which occur with probability 1 in the given document class. The penalty and boosting values used in the algorithm were determined experimentally. The score we obtain is the probability of the given document belonging to a particular document class. Surprisingly, this simple approach gives us much better accuracy rates than Naïve Bayes, which we thought would be a good option since it is widely used in classification tasks like spam filtering [3]. We have decent accuracy rates with this simple, naïve algorithm; one future task could be to develop a better classifier.

Table 4.3 shows the probability distribution for the adjectives in a lyrics snippet in Figure 4.1. The algorithm correctly identifies the document to be lyrics.

Word	P(word given lyrics)	P(word given articles)	P(word given poetry)
even	0.9003	0.8863	0.9195
true	0.0406	0.0462	0.1087
deep	0.3549	0.1043	0.2672
impossible	0.9782	0.8299	0.9565
bad	0.1978	0.0776	0.0609
Total	2.4718	1.9443	2.3128

Table 4.3: The probability distributions for the adjectives in the lyrics snippet in Figure 4.1. The overall classification for the snippet is that it is lyrics.

4.5 Results

First, we look at the classification accuracies among lyrics, articles and poems obtained by our classifier. We show that the adjectives used in lyrics are much more rhymable than the ones used in poems but they do not differ significantly in their semantic orientations. Furthermore, our algorithm is able to identify poetic lyricists from non-poetic ones using the word distributions calculated in an earlier section. We also compare adjectives for a given concepts which are more likely to be used in lyrics rather than poetry and vice versa.

4.5.1 Document Classification

Our test set consists of the text of 100 examples from each of our three categories. Using our algorithm with the adjective distributions we get an accuracy of 67% for lyrics, 80% for articles and 57% for poems.

The confusion matrix showing the performance of our algorithm is in Table 4.4. We find that we can detect articles with the highest accuracy, this might be because of the enormous size of the article training set which consisted of all English Wikipedia articles. A slightly more number of articles get misclassified as lyrics than poetry. A large number of misclassified poems get classified as articles rather than lyrics, but most misclassified lyrics get classified as poems. Typically poetry is written for a more sophisticated and educated audience; newspaper and wikipedia articles (our article data set) targets similar audience. Hence it is plausible that a majority of poetry gets misclassified as articles, but a majority of articles getting misclassified as lyrics rather than poetry is surprising.

For example, the first step of running the document classification algorithm on the song snippet in Figure 4.1 will consist of detecting all the adjectives. We then proceed to get the synonyms of all the adjectives which have previously been detected using the method described in Section 4.4.1. We then proceed to calculate the score of the document being lyrics, poetry or article using the probability distributions for each of the document class. The probability distribution of adjectives for each document class is calculated using the method described in Section 4.4.2. The document class with the highest score is the predicted class of the given document.

4.5.2 Adjective Usage in Lyrics versus Poems

Poetry is written against a silent background while lyrics are often written keeping the melody, rhythm, instrumentation, the quality of the singers voice and other qualities of the recording in mind [70]. Furthermore, unlike most poetry, lyrics include repeated lines [70]. This led us to believe the adjectives which were more likely to be used in lyrics rather than poetry would be more rhymable.

Rhymezone [71] is a website which maintains a list of word a given word rhymes with. We counted the number of words an adjective in our lyrics and poetry list rhymes with from their website. The values are tabulated in Table 4.5. From the values in Table 4.5, we can clearly see that the adjectives which are more likely to be used in lyrics to be much more rhymable than the adjectives which are more likely to be used in poetry. For example, “happy”, which is more likely to be used in poetry rather than in lyrics, rhymes

with 10 other words. Its synonym “elated” is more likely to be used in lyrics rather than in poetry and rhymes with 56 other words. This is quite plausible, since a good song is made by its hooks, and there are several different kinds of hook. One of them is catchy rhyme, which can contribute to the memorizability of songs [89].

	Predicted		
Actual	Lyrics	Articles	Poems
Lyrics	67	11	22
Articles	11	89	6
Poems	10	33	57

Table 4.4: The confusion matrix for document classification. Many lyrics are categorized as poems, and many poems as articles.

	Lyrics	Poetry
Mean	33.2	22.9
Median	11	5
Standard Deviation	58.37	46.51
25 th percentile	2	0
75 th percentile	38	24

Table 4.5: Statistical values for the number of words an adjective rhymes with. The two-tailed P value is less than 0.0001.

	Lyrics	Poetry
Mean	-0.050	-0.053
Median	0.0	0.0
Standard Deviation	0.328	0.334
25 th percentile	-0.27	-0.27
75 th percentile	0.13	0.13

Table 4.6: Statistical values for the semantic orientation of adjectives used in lyrics and poetry. The two-tailed P value equals 0.8072.

We were also interested in finding if the adjectives used in lyrics and poetry differed significantly in their semantic orientations. SentiWordNet assigns to each synset of WordNet two sentiment scores: positive sentiment score and negative sentiment score. For each

adjective in our lyrics and poetry data set we calculated the net semantic orientation by calculating the difference between the positive and negative sentiment score for all of the synsets it belonged to. For example, happy belongs to four synsets and has a net semantic orientation of 0.5625. We calculated the semantic orientations, which take a value between -1 and +1, using SentiWordNet, of all the adjectives in the lyrics and poetry list, the values are in Table 4.6. They show no difference between adjectives in poetry and those in lyrics.

4.5.3 Poetic vs non-Poetic Lyricists

We were curious if our method would help us detect poetic lyricists from non-poetic ones since the choice of words of poetic lyricists might be influenced by their poetic style. There are lyricists like Bob Dylan [61], Ani DiFranco [25], and Stephen Sondheim [82, 81], whose lyrics are considered to be poetic, or indeed, who are published poets in some cases. The lyrics of such poetic lyricists possibly could be structurally more constrained than a majority of the lyrics or might adhere to a particular meter and style.

While selecting the poetic lyricists we ensured that popular articles supported our claim or by going to their Wikipedia page and ensuring that they were poets along with being lyricists and hence the influence of poetic forms on their lyrics. The list of poetic lyricists and the justification behind choosing them is in Table 4.2.

Our algorithm consistently misclassifies a large fraction of the lyrics of such poetic lyricists as poetry while the percentage of misclassified lyrics as poetry for the non-poetic lyricists is significantly much less. These values for poetic and non-poetic lyricists are tabulated in Table 4.7 and Table 4.8 respectively.

Poetic Lyricists	Number of Lyrics	% of lyrics misclassified as poetry
Bob Dylan	377	42%
Ed Sheeran	93	50%
Ani Di Franco	213	29%
Annie Lennox	88	32%
Bill Callahan	6	34%
Bruce Springsteen	513	29%
Stephen Sondheim	21	40%
Morrissey	227	39%
Totals	1538	36%

Table 4.7: Percentage of misclassified lyrics as poetry for poetic lyricists.

Non-Poetic Lyricists	Number of Lyrics	% of lyrics misclassified as poetry
Bryan Adams	232	14%
Michael Jackson	321	22%
Drake	144	7%
Backstreet Boys	248	23%
Radiohead	240	26%
Stevie Wonder	300	17%
Led Zeppelin	118	8%
Kesha	146	18%
Totals	1749	17%

Table 4.8: Percentage of misclassified lyrics as poetry for non-poetic lyricists.

From the values in Table 4.7 and Table 4.8 we see that there is a clear separation between the misclassification rate between poetic and non-poetic lyricists. The maximum misclassification rate for the non-poetic lyricists, 26% is less than the minimum misclassification rate for poetic lyricists, 29%. Furthermore the difference in average misclassification rate between the two groups of lyricists is 19%. Hence our simple algorithm can accurately identify poetic lyricists from non-poetic ones, based only on adjective usage: with enough examples, poetic lyricists are those that routinely have their lyrics classified as poetry, not lyrics.

4.5.4 Concept representation in Lyrics vs Poetry

We compare adjective uses for common concepts. To represent physical beauty we are more likely to use words like “sexy” and “hot” in lyrics but “gorgeous” and “handsome” in poetry. For 20 of these, results are tabulated in Table 4.9. The difference could possibly be because unlike lyrics, which are written for the masses, poetry is generally written for people who are interested in literature [70].

4.6 Conclusion

We have developed a method to detect the genre of a document based on the probability distribution of synonymous adjectives. Our key finding is that the choice of synonym for even a small number of adjectives is sufficient to reliably identify the genre of documents.

In accordance with our hypothesis, we show that there exist differences in the kind of adjectives used in different genres of writing. We calculate the probability distribution of synonymous adjectives over the three kinds of documents and using this distribution and a simple algorithm, we are able to distinguish among lyrics, poetry and article with an accuracy of 67%, 57% and 80% respectively. Using our algorithm we show that we can discern poetic lyricists like Bob Dylan and Stephen Sondheim from non-poetic ones like Bryan Adams and Kesha. Our algorithm consistently misclassifies a majority of the lyrics of such poetic lyricists as poetry while the percentage of misclassified lyrics as poetry for the non-poetic lyricists is significantly lower.

Lyrics	Poetry
proud, arrogant, cocky	haughty, imperious
sexy, hot, beautiful, cute	gorgeous, handsome
merry, ecstatic, elated	happy, blissful, joyous
heartbroken, brokenhearted	sad, sorrowful, dismal
real	genuine
smart	wise, intelligent
bad, shady	lousy, immoral, dishonest
mad, outrageous	wrathful, furious
royal	noble, aristocratic, regal
pissed	angry, bitter
greedy	selfish
cheesy	poor, worthless
lethal, dangerous, fatal	mortal, harmful, destructive
afraid, nervous	frightened, cowardly, timid
jealous	envious, covetous
lax, sloppy	lenient, indifferent
weak, fragile	feeble, powerless
black	ebon
naïve, ignorant	innocent, guileless, callow
corny	dull, stale

Table 4.9: For twenty different concepts, we compare adjectives which are more likely to be used in lyrics rather than poetry and vice versa.

The algorithm developed has many practical applications in Music Information Retrieval (MIR). As we have shown, we can analyze documents, analyze how lyrical, poetic

or article-like a document is. For lyricists or poets we can come up with alternate better adjectives to make a document fit its genre better. Using the word distributions we can come up with a better measure of distance between documents where the weights are assigned to a word depending on its probability of usage in a particular type of document. And, of course, our work here can be extended to different genres of writings like prose or fiction.

Chapter 5

Can Song Lyrics Predict Hits?

5.1 Introduction

Can we predict if a song will be a hit before it is even released? This music information retrieval task, sometimes called hit song science [63], has traditionally been seen as extremely hard [68]. Solving it would be of immense use to music label companies. They want to invest resources in songs likely to become hits and give a good return on their investment, rather than publicizing songs set to be flops. Successful hit detection might also identify talented music artists whose songs otherwise would not have received enough airplay time. Most previous work in hit song detection has been of modest success, and has typically focused on audio aspects of a song recording [63, 68], though Dhanaraj and Logan [24] used both text and audio features in their experiments.

Our focus in this thesis is on studying lyrics as a component of the artistic creation in a song. Since lyrics are typically set in verses and choruses, we analyze the structure of these elements, with a focus on rhyme, meter and syllable content. Lyrics contain much of the emotional content of a typical popular song [2], and are also a much smaller input set (typically a few kilobytes for a song) than the megabytes to analyze in a recording of the audio of the song. Lyrics also contribute to the memorizability of songs [89], and offer songwriters the opportunity to show off their creativity in an easily noticed fashion, such as through clever wordplay or rhyme patterns. Finally, behavioral and neuropsychological research has shown that individuals process lyrics and tune separately while listening to songs [37].

We make use of song lyrics to build our models, though in later experiments we added 14 audio features from Echo Nest [62] to analyze the effectiveness of incorporating the

audio recording into prediction of hits. We use the complete set of 24 rhyme and syllable features of the Rhyme Analyzer [42], and add seven new meter features identifying the fraction of lines written in a particular meter. The description of our 31 lyrics features is in Table 5.4.

As is often true with music information retrieval tasks, the core question in our work is the separation of types of recordings that are hard to define: what is a hit, and what is a flop? We define hits as songs that made it to the Billboard Year-End Hot 100 singles chart in the years 2008-2013. We select recent hits since pop music evolves over time: a 1960's-era Beatles hit, which no doubt a lot of people still listen to, might be a flop in 2015. By contrast, it might be difficult to come to a consensus on the definition of a flop, so we use several different definitions of flops, ranging from a very broad one to extremely restricted ones.

We use standard machine-learning algorithms, such as weighted support vector machine [19] and Bayesian network classifiers from Weka [35], with 10-fold cross validation. The SVM slightly outperforms the Bayesian network, but we focus our presentation on the Bayes net classifiers, for simplicity of presentation. Surprisingly, the Bayesian networks we obtain from Weka are naïve Bayes: the effect of one feature is independent of another.

Our major results are twofold. First, the simple classifiers we build from lyric features are surprisingly effective at separating flops and hits. And second, in all cases where one of our rhyme or meter features is helpful in predicting whether a song is a hit or not, we find that the more complex a song's rhyme or meter, the more likely it is a hit.

A limitation of focusing on lyrics is that it prevents us from distinguishing good and bad covers of the same song by different artists singing the same words. Similarly, a song might become a hit on the basis of great instrumental work or a terrific video: our methods cannot be expected to succeed in these cases either. However, a key result of our work is that we can distinguish clever lyrics from less clever ones, and that this separation allows us to identify at least one aspect of high-quality, or at least successful, songwriting.

The results of this Chapter has previously been presented as a late breaking demo at ISMIR 2014 [77] and at CMMR 2015 [79]. This chapter quotes extensively from these publications.

5.2 Related Work

Several authors have previously attempted to predict hit songs, largely using audio features. Dhanaraj and Logan [24] use both text and audio features individually and together to

predict hit songs. They take songs which made it to the number 1 position in the United States, United Kingdom, or Australia from January 1956 to April 2004 as hits. They do not describe the songs that they consider flops. They learn the most prominent sounds and topics of each song (using textual analysis), and conclude that the text features are slightly more useful than the audio features; combining both of them together does not produce significant improvements. They obtain an average area under ROC curve of 0.66 using the audio features, while using the text features, or combining both types of features, gives an average area under ROC curve of 0.68 and 0.69 respectively. They focus on the semantic content of the words, learning the topic every song belongs to in their lyrics collection. A key limitation of their work is that they used features designed for prose rather than ones designed for verse. Ni et al. [63] use audio features to discern the top 5 hits from the other top 30-40 hits using a shifting perceptron. They achieve classification accuracy of slightly more than 50% across all the decades from 1960-2010. Pachet and Roy [68] attempt to use spectral features like chroma, spectral centroid, skewness, and manually entered labels, to learn a label of low, medium or high popularity using a support vector machine with boosting. They conclude that using their features, it is not possible to gauge the popularity of a song.

Fan and Casey [31] used a set of ten common audio features such as energy, loudness and danceability with a time weighted linear regression model and a support vector machine model to predict Chinese and UK pop hits from a data set of 347 Chinese and 405 English songs. They conclude that Chinese hit song prediction is easier than British hit song prediction and show that the audio feature characteristics of Chinese hit songs are significantly different from those of UK hit songs. They obtain an error rate of slightly more than 41% and 39% for English and Chinese songs respectively on balanced data. Herremans et al. [38] used audio features to discern Top 10 dance song hits from songs with lower listed position. They obtained the best results with logistic regression closely followed by naïve Bayes classifier.

Bischoff et al. [13] exploit social annotations and interactions in Last.fm and the relationships between tracks, artists and albums to predict hits. Since these social tags incorporate lots of information about why hits are hits, they are clearly of a different sort than those that are based on the primary creative work only.

Only one group has previously focused on the properties of lyrics that distinguish them as not being prose: Smith et al. [80] make use of TF-IDF weighting to find typical phrases and rhyme pairs in song lyrics and conclude that typical number one hits, on average, are more clichéd.

Though our audio features are similar to the ones used by previous work, we are unaware

of any previous work which uses meter and syllable features for hit detection. Unlike Smith et al. [80], which concentrates on cliched rhymes, we consider all rhyming pairs in the lyrics, including imperfect and line-internal rhymes.

Unfortunately, it is difficult to compare the results from our work with those from previous works: these works do not provide a confusion matrix, nor do they provide easily interpretable results. There are some key differences and enhancements between our work and previous works: the use of unique lyrical features, exploring different definitions of flops, and providing complete results of our experiments. We are unaware of any previous work that explored the consequence of using different definitions of flops, but this is key, as each definition has its own pros and cons and warrants attention. Unlike previous work with songs spanning a few decades [24], we select hits from the shorter 2008-2013 interval, as music taste evolves over time. Also, our data sets are available at a website, www.cs.uwaterloo.ca/~browndg/CMMR15data.

5.3 Rhymes

A rhyme is a repetition of similar sounds or the same sound in two or more words, most often in the final syllables of lines in poetry and lyrics. They can be further divided into degree and manner of phonetic similarity into perfect and general rhymes. Perfect rhymes have their final stressed vowel and the following sound identical, as in sight and flight, deign and gain, madness and sadness [39]. General rhyme can refer to the rhyme between various kinds of phonetic similarity between words, as in “wing” and “caring”, “bend” and “ending”, “shake” and “hate”. In this thesis the focus is on imperfect rhymes. Rhymes can also be classified according to their position in the verse. We here focus on internal rhymes. The examples that follow come from Hirjee’s Master’s thesis [39].

5.3.1 Imperfect Rhymes

Imperfect rhyme is a type of rhyme formed by words with similar but not identical sounds. In most instances, either the vowel segments are different while the consonants are identical, or vice versa. Examples include:

When have I last looked **on**
The round green eyes and the long wavering bodies
Of the dark leopards of the **moon?** [96]

where “on” and “moon” are imperfect rhymes.

5.3.2 Internal Rhymes

Internal rhymes occurs when a word or phrase in the interior of a line rhymes with a word or phrase at the end of a line, or within a different line. They are of the following types: chain rhymes, compound rhymes, and bridge rhymes.

Chain rhymes are consecutive words or phrases in which each rhymes with the previous, as in:

New York **City gritty** committee **pity** the fool that
Act shitty in the midst of the calm the witty [60]

where “city”, “gritty”, “committee”, and “pity” participate in a chain since they all rhyme and follow each other contiguously

Compound rhymes are formed when two pairs of line internal rhymes overlap within a single line, as in:

Yo, I stick around like hockey, now what the puck
Cooler than **fuck**, *maneuver* like *Vancouver* **Canucks** [60]

where “maneuver” and “Vancouver” are found between “fuck” and “Canucks.”

Bridge rhymes are internal rhymes spanning two lines where both the members are not line final, as in:

Still I be packin **agilities** unseen
Forreal-a my killin **abilities** unclean facilities. [60]

where “agilities” and “abilities” are bridge rhymes.

Link rhymes are internal rhymes spanning two lines where the first word or phrase is line-final, as in:

How I made it you salivated over my **calibrated**
Raps that **validated** my ghetto credibility [60]

where “calibrated” and “validated” are link rhymes.

5.4 Data Definition

A specific focus of our project has been to rigorously define the two groups of songs that we wish to separate. We define hits as songs which made it to the Billboard Year-End Hot 100 singles chart between 2008 and 2013, eliminating duplicate songs repeated across two years. This leaves 492 hits.

Far more challenging has been the definition of “flop.” With no “flops chart” it is difficult to come to a consensus on this concept. Previous authors [63] have used the songs at the lower end of the top 100 year end charts as flops, but we believe that those songs are not flops since very popular songs of genres with relatively few listeners can end up in those positions. Hence, we conduct experiments on four different definitions of flops, ranging from broad to very narrow. The number of hits and flops in our data set for the four definitions of flops is in Table 5.1.

	Hits	Flops	Total
Definition 1	492	6323	6815
Definition 2	492	1131	1623
Definition 3	92	234	326
Definition 4	492	765	1257

Table 5.1: The number of hits and flops in our data set using the four definitions of flops.

For our first exploration, we start by defining a set of 57 artists who have had massive hit songs in the 2008-2013 period. These “hit artists” include household American, Canadian and British names like Lady Gaga, Justin Bieber, and Adele. Fifteen of them are listed in Table 5.2. In this framework, a flop is a song released by a hit artist that did not reach the definition of “hit”. However, many artists include songs on their albums that cannot be expected to be huge hits, so this definition of flop is broad: it does not take into account songs which could have become a hit had they received enough airplay time.

Artist	Years active
Lady Gaga	2001 - Present
Justin Bieber	2008 - Present
Adele	2008 - Present
Maroon 5	1994 - Present
One Direction	2010 - Present
Flo Rida	2006 - Present
Kesha	2005 - Present
Macklemore	2000 - Present
Nicki Minaj	2004 - Present
Rihanna	2005 - Present
Taylor Swift	2004 - Present
Calvin Harris	1999 - Present
Pitbull	2001 - Present
Eminem	1988 - Present
Jason Derulo	2006 - Present

Table 5.2: Fifteen of the artists in our data set.

In our second definition, we define flops as songs which made it to the Billboard weekly Hot 100 chart between 2008 and 2013 but did not make it to the Billboard Year-End Hot 100 singles chart. It might be argued that any song ever on the weekly Top 100 is not a flop, but being on the weekly chart does show that it received adequate airplay time, and its promoters might have hoped it would be a major hit, not just a brief flash in a pan.

For the third definition, we take flops to be songs which made it to the Billboard Year-End chart in 2013 for any of thirteen different genres (pop, rock, *etc.*) but did not make it to the Billboard Year-End Hot 100 singles chart for that year. This is an extremely restrictive definition. Popular songs of relatively less-heard genres, which might not be expected to make it to the year end Hot 100 charts, can be wrongly considered to be flops by the definition. For example, in 2013 the country song, “Better Dig Two” was at the 13th position in the year-end country chart but did not make it to the year-end genre-independent (Hot 100) chart. Our third definition declares this song a flop, though it has over 10 million views on YouTube.

Our final proposed definition is that flops are songs by hit artists that were released as singles, but did not make it to the Billboard Year-End Hot 100 singles chart between 2008 and 2013. Arguably, this is the best definition of flops since music labels spend a lot of resources in promoting singles, and such songs do get airplay: the only reason they do not

make it to the year end chart is because of negative response of listeners. On the other hand, singles are much less relevant now than they once were [75].

For all songs, we obtained lyrics from a free online lyrics repository [57]. On manual inspection of the lyrics of flops we observe that the stored lyrics of flops that are shorter than thirty lines are very noisy on lyrics websites, with misspellings, errors or repetitions of meaningless syllables like “lalala”. It is hard to automatically predict rhyme features on messy lyrics. Thus, we only study songs with at least thirty lines of lyrics. Since almost all the hits were greater than thirty lines we did not eliminate any based on their lengths, though many short hits were hard to classify. We downloaded Billboard charts from Billboard’s website [10], while the list of single releases of artists were obtained from the artists’ discography pages on Wikipedia [92].

5.5 Method

We use the complete set of 24 rhyme and syllable features of the Rhyme Analyzer [42], a tool developed to analyze hip-hop lyrics, and that finds rhymes, including imperfect rhymes (like “time” rhyming with “line”) and internal rhymes (where both elements of a rhyming pair are not in line-final position), and calculates syllable features. The description of these internal rhymes is in Section 5.3. We refer the reader to Hirjee and Brown [43] for more information about these features.

We use a total of 31 lyrics features, which are defined in Table 5.4. Lyrics, unlike prose, adhere to certain structure, and these features can both separate lyrics from prose and may identify high-quality songmaking craftsmanship. We add seven new meter features identifying the fraction of lines written in iambic, trochaic, spondaic, anapestic, dactylic, amphibrachic and pyrrhic meter, using the CMU Pronunciation Dictionary [28] to transcribe lyrics to a sequence of phonemes with indicated stress. In this framework, spondaic meter indicates a line entirely of stressed syllables, and pyrrhic means line of entirely unstressed syllables. The other meters have patterns with stressed and unstressed syllables. The stress pattern of the meter features we use can be found in Table 5.5. We use a total of 31 lyric features, the definition of which can be found in Table 5.4. Lyrics, unlike prose, is expected to adhere to certain structure and these features separate lyrics from prose and can be considered to be proxy for craftsmanship. We use different definitions of flops, as described in the previous section, and include no flops shorter than 30 lines.

Feature	Definition
---------	------------

Syllables per Line	Average number of syllables per line
Syllables per Word	Average word length in syllables
Syllable Variation	Standard deviation of line lengths in syllables
Novel Word Proportion	Average percentage of words in the second line in a pair not appearing in the first
Rhymes per Line	Average number of detected rhymes per line
Rhymes per Syllable	Average number of detected rhymes per syllable
Rhyme Density	Total number of rhymed syllables divided by total number syllables
End Pairs per Line	Percentage of lines ending with a line-final rhyme
End Pairs Grown	Percentage of rhyming couplets in which the second line is more than 15% longer in syllables than the first
End Pairs Shrunk	Percentage of rhyming couplets in which the second line is more than 15% shorter in syllables than the first
End Pairs Even	Percentage of rhyming couplets neither grown or shrunk
Average End Score	Average similarity score of line-final rhymes
Average End Syl Score	Average similarity score per syllable in line final rhymes
Singles per Rhyme	Percentage of rhymes being one syllable long
Doubles per Rhyme	Percentage of rhymes being two syllables long
Triples per Rhyme	Percentage of rhymes being three syllables long
Quads per Rhyme	Percentage of rhymes being four syllables long
Longs per Rhyme	Percentage of rhymes being longer than four syllables
Perfect Rhymes	Percentage of rhymes with identical vowels and codas
Line Internals per Line	Number of rhymes with both parts falling in the same line divided by total number of lines
Links per Line	Average number of link rhymes per line
Bridges per Line	Average number of bridge rhymes per line
Compounds per Line	Average number of compound rhymes per line
Chaining per Line	Total number of words or phrases involved in chain rhymes divided by total number of lines
Iambic proportion	Percentage of lines in iambic meter
Trochaic proportion	Percentage of line in trochaic meter
Spondaic proportion	Percentage of line in spondaic meter
Anapestic proportion	Percentage of line in anapestic meter
Dactylic proportion	Percentage of line in dactylic meter
Amphibrachic proportion	Percentage of line in amphibrachic meter

Pyrrhic proportion	Percentage of line in pyrrhic meter
--------------------	-------------------------------------

Table 5.4: The list of lyric features used by our algorithm. Singles per rhyme is the only feature that argues against lyrical complexity and is more relevant in Chapter 6.

Most previous works in hit detection [68, 63, 38] have used audio features to discern hits from flops. Some, like Dhanaraj and Logan [24], used text features or combine both audio and text features to predict hits. Though our main focus was on analyzing the use of rhyme, meter and syllable features for hit detection, we were curious about the usefulness of audio features in predicting hits. We added 14 audio features: danceability, loudness, energy, mode, tempo, and the mean, median and standard deviation of the timbre, pitch and beat duration vectors. These features are often used in MIR work, as they have been pre-computed for the Million Song Database [8]. We obtained these audio features via the Echo Nest’s APIs [62]. We discarded the songs for which we could not find the audio features from Echo Nest [62]; this failure might be because of Echo Nest not having the data or due to incorrect song or artist spelling. We are left with 476 hits and 3179 flops on which we run the experiments using just the audio features and combining both audio and lyrics features together.

Our experiments have unbalanced data sets; since there are many more flops than hits for three of our four definitions of “flop”. Our largest data set consists of 492 hits and 6323 flops. We used weighted-cost SVMs, in LIBSVM [19], which assign different misclassification cost to instances depending on the class they belong to. Tuning the misclassification costs, we can adjust the number of true and false positives: large and small values of misclassification cost give trivial classifiers, while intermediate costs trade false negatives for false positives at different ratios. We also used the Bayesian network module from Weka [35] with ten-fold cross validation, and we report the confusion matrices for the network that maximizes data likelihood. Similarly we use Weka and ten-fold cross validation for weighted-cost SVMs and run the experiments for different misclassification costs, selecting weights so that recall is close to 50%. In what follows, we focus on the Bayesian networks because they are easier to explain.

Meter	Stress pattern
Iambic	Unstressed + Stressed
Trochaic	Stressed + Unstressed
Spondaic	Stressed + Stressed
Anapestic	Unstressed + Unstressed + Stressed
Dactylic	Stressed + Unstressed + Unstressed
Amphibrachic	Unstressed + Stressed + Unstressed
Pyrrhic	Unstressed + Unstressed

Table 5.5: Stress pattern in different meters.

5.6 Results

Lyrics features can quite effectively separate hits and flops. For our broadest definition of flops, we can correctly detect around half of the hits and misclassify just 12.8% of flops as hits. A summary of our results are in Tables 5.6 and 5.7.

Filters produced from lyrics features significantly outperform the best filter we can build for audio features, for all definitions of flops. Combining the audio and lyrics features gives us the best results, but they are not much better than the ones obtained solely using the lyrics features. We perform experiments for all the definitions of flops as defined in Section 5.4. From our Bayesian network we find that our most important features are rhymes per line, rhyme density, end pairs shrunk, link and line internal rhymes per line, and the audio feature of loudness. These features indicate that songs with lots of rhymes, in particular complicated ones, are more likely to end up becoming hit. We argue that complex rhyme and meter is a detectable property of lyrics that indicates quality songmaking and artisanship and allows artists to become successful.

5.6.1 The Broadest Definition of Flop

In our first experiment, hits are songs which made it to the Billboard Year-End Hot 100 singles chart between the years 2008-2013 while flops are songs by hit artists which did not make it to the year-end chart. The results using a weighted-cost SVM and Bayes net is shown in Table 5.6 and 5.7 respectively. We are able to correctly detect around half of the hits and misclassify just 12.8% of flops as hits. A weighted-cost SVM outperforms Bayesian

networks in detecting hits with better precision and recall values. Since the data sets are unbalanced, we assign different penalties to misclassified instances of different classes. A weight of 8 implies that it is 8 times more expensive to misclassify a hit (a false negative) than a flop (a false positive). We obtain the best result when using a weight of 8, the confusion matrix for which is in Table 5.6 and the receiver operating characteristic curve plot is the left curve in Figure 5.1. Here, we have tuned the weight parameter to see what the precision is when the recall is close to 50%. We focus on confusion matrices and not area under curve because a high-AUC classifier is not always better than a low-AUC one [32] and the matrices give a more complete picture.

	Definition 1	Definition 2	Definition 3	Definition 4
# correctly classified hits	248	223	45	250
# misclassified hits	244	269	47	242
# correctly classified flops	5514	756	133	493
# misclassified flops	809	375	101	272
precision(Hits)	0.253	0.373	0.308	0.479
recall(Hits)	0.504	0.453	0.489	0.504
F-score(Hits)	0.337	0.409	0.378	0.491

Table 5.6: The results we obtain using a weighted-cost SVM. The weights are chosen to keep the recall close to 50%.

	Definition 1	Definition 2	Definition 3	Definition 4
# correctly classified hits	222	69	0	124
# misclassified hits	270	423	92	368
# correctly classified flops	5510	1032	0	638
# misclassified flops	813	99	234	127
precision(Hits)	0.214	0.411	0.0	0.494
recall(Hits)	0.451	0.14	0.0	0.252
F-score(Hits)	0.290	0.209	0.0	0.333

Table 5.7: The results we obtain using a Bayesian network.

We added 14 audio features and repeated the experiment on the same 6815 songs as used above, (without the audio features for 3160 songs), using just the lyrics features, just the audio features and both lyrics and audio. We see that lyrics features are significantly better than audio features in predicting hits, and that adding audio features only slightly

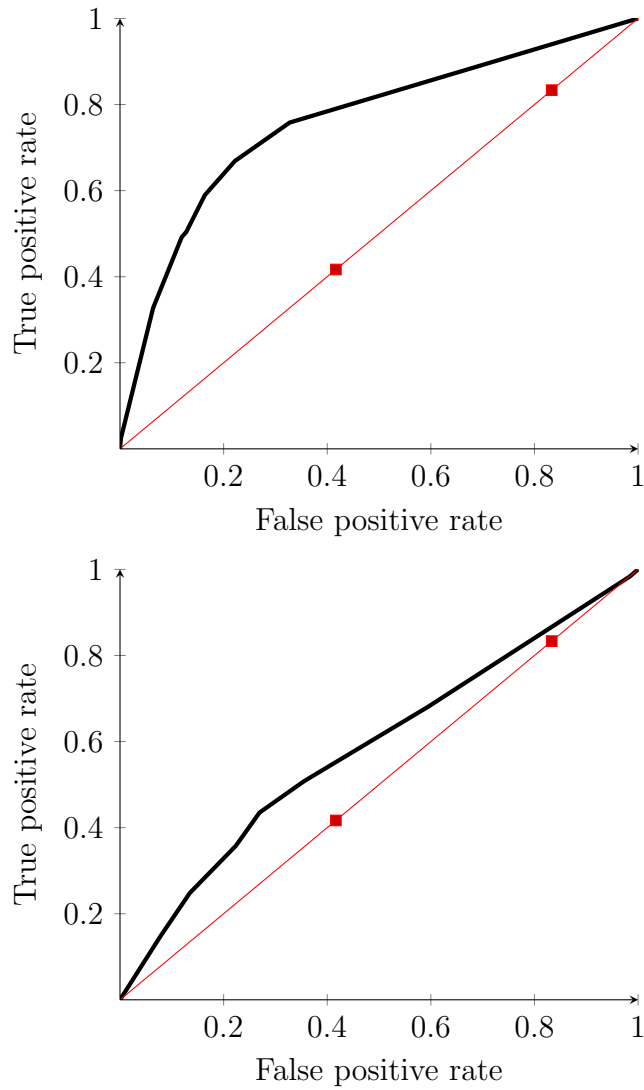


Figure 5.1: The receiver operating characteristic curve obtained when using the first and second definition of flops respectively. The areas under the ROC curves are 0.688 and 0.573 respectively.

improved the results. The results obtained using a Bayesian network are in Table 5.8 and the ROC curve is depicted in Figure 5.2. The confusion matrices obtained using both the audio and lyrics features and a weighted-cost SVM are in Table 5.9.

We observe that the performance of our algorithm increases considerably as the length of the lyrics increases. We believe that this is because the probability of lyrics being noisy decreases as its length increases; we verified this by manually inspecting flops. Repeating the above experiment with flops which are at least fifty lines long and using a Bayesian network, we obtain the confusion matrix shown in Table 5.10. As most hit lyrics are lengthy and relatively noise free we do not eliminate them based on their line count. Our approach works especially well for relatively noise-free, lengthy lyrics.

	Lyrics	Audio	Audio+Lyrics
# correctly classified hits	218	105	235
# misclassified hits	259	371	241
# correctly classified flops	2656	2818	2680
# misclassified flops	523	361	499
precision(hits)	0.294	0.225	0.318
recall(hits)	0.457	0.221	0.491
F-score(Hits)	0.358	0.223	0.386

Table 5.8: Lyrics features outperform the audio features in discerning hits from flops.

Predicted Value		
True Value	Hits	Flops
Hits	237	239
Flops	745	2434
Precision(Hits) = 0.241		
Recall(Hits) = 0.498		
F-Score(Hits) = 0.323		

Weight 3.5 SVM, audio+lyrics

Table 5.9: Surprisingly, the naïve Bayesian network gives us better result than weighted-cost SVM when using both audio and lyrics features.

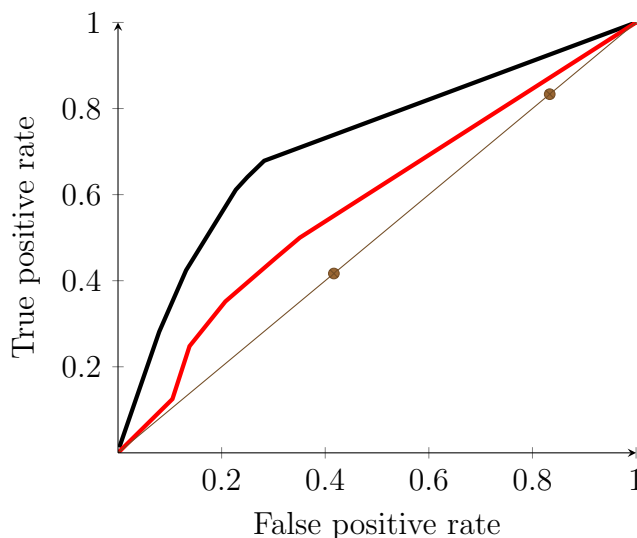


Figure 5.2: The ROC curve obtained when using the first definition of flop and a weighted-cost SVM. The black and the red curves are obtained using the lyrics and audio features respectively. The AUC using the lyrics and audio features is 0.692 and 0.572 respectively.

True Value	Predicted Value	
	Hits	Flops
Hits	218	274
Flops	305	1780
Precision(Hits) = 0.416		
Recall(Hits) = 0.443		
F-Score(Hits) = 0.429		

Bayesian network, Lyrics only

Table 5.10: We see a noticable improvement in performance with lyrics longer than 50 lines, which are more accurate than the shorter ones. Compare to Table 5.7

5.6.2 The “Flash in a Pan” Definition of Flop

As noted earlier, the first definition of flop is broad as we classify songs with no airplay time as flops. In this experiment, hits are songs which made it to the Billboard Year-End Hot 100 singles chart between the years 2008-2013, while flops are songs which made it to the Billboard weekly Hot 100 chart between 2008 and 2013 but never rose to the Billboard

Year-End Hot 100 singles chart. Inclusion in the weekly chart indicates that a song received adequate air play time and had the potential to be a hit. The results using a weighted-cost SVM and Bayes net is shown in Table 5.6 and 5.7 respectively, choosing a weight for the SVM that gives a recall $\approx 50\%$, and the ROC plot is the right curve in Figure 5.1. Despite this new definition being more restricted than the first one, we see better accuracies than in the first experiment. We correctly identify almost half of the hits while misclassifying 33.16% of flops as hits using the SVM.

5.6.3 The “Hit on One Chart” Definition of Flop

In this experiment we take hits to be songs which made it to the Billboard Year-End Hot 100 singles chart in 2013 while flops are songs which made it to the Billboard year end chart for thirteen different genres: pop, gospel Christian, dance club, dance electronica, rap, R&B, hip-hop, alternative, rock, country, adult pop and adult contemporary, in 2013 but did not make it to the 2013 Billboard Year-End Hot 100 singles chart. The results using a weighted-cost SVM and Bayes net is shown in Table 5.6 and 5.7 respectively. Surprisingly, using Bayesian network we obtain a trivial classifier, which may be because of the small data set. The weighted-cost SVM does a much better job and is tuned so that recall is close to 50%.

Any song, irrespective of its genre, which makes it to a year-end genre specific chart, is probably a hit, while the songs which make it to the genre independent year-end chart are “mega-hits”. This problem of differentiating “hits” from “mega-hits” is extremely difficult and we are successfully able to identify around half of the hits and misclassify 43.16% of flops as hits. This is probably the most challenging task we consider, and our results are not very strong.

5.6.4 The “Not-Hit Single” Definition of Flop

The previous definition of a flop is extremely restrictive, as popular songs of relatively less-heard genres, which might not be expected to make it to the Hot 100 year end charts, are not really flops. A flop should be a song which receives airplay time but never becomes popular among the masses. In our final exploration we define flops to be songs by one of our identified hit artists that were released as singles, but did not make it to the Billboard Year-End Hot 100 singles chart between 2008 and 2013. Arguably, this is the best definition of flops since music labels spend a lot of resources in promoting their singles, and such songs do get airplay. The results using a weighted-cost SVM and Bayes net is shown in Table

5.6 and 5.7 respectively, again choosing a weight for the SVM that gives us recall $\approx 50\%$. We correctly identify half of the hits while misclassifying only 35.56% of flops as hits.

5.7 What makes a song hit?

Feature	% of hits	% of flops
Rhymes per line ≥ 3.016	18.08	6.17
Rhyme density ≥ 0.594	15.44	5.19
End pairs shrunk ≥ 0.735	13.21	3.33
Link rhymes per line ≥ 0.527	8.73	2.29
Line internal rhymes per line ≥ 1.618	17.68	5.72
Loudness ≤ 12.909	24.8	10.0

Table 5.11: The most important features and their values for hit detection and the percentage of hits and flops falling in that range.

Artist	Correctly classified hits	Misclassified hits
Maroon 5	Misery, Daylight, One More Night	Payphone, Moves Like Jagger
Adele	Set Fire to the Rain, Someone Like You, Rumour Has It	Rolling in the Deep
Lady Gaga	Bad Romance, Applause, Just Dance	Born This Way, Paparazzi, Telephone
Leona Lewis	Bleeding Love	Better in Time

Table 5.12: The outcome of our algorithm on hits in our data set using the first definition of flops for four popular artists.

Perhaps the most surprising, and indeed heartening, result from our experiments is that rhyme, meter and lyrics matter in hit detection and that complexity is connected to being a hit. By contrast, none of our audio features allow for an investigation of these parameters for the musical part of a song. Surprisingly, loudness is the only important audio feature: very loud songs are more likely to end up as flops. Table 5.11 lists some of the most important features, their boundary values for the hit detection, and the percentage of hits and flops falling in that range using the Bayesian network coming from our first definition

of flops. For example, “One More Night,” a very popular song by Maroon 5, is correctly identified as a hit because of the presence of frequent complicated rhymes. “Payphone,” another popular Maroon 5 song is misclassified as a flop due its comparative simplicity. Similarly, extremely popular songs like “Rolling in the Deep,” “Born this Way,” *etc.* are misclassified as flops due to their fewer, rhymes. These songs may have hit for other reasons, of course. Table 5.12 lists the outcome of our algorithm on hits in our data set for four popular artists.

We do not claim that the presence of these features make a hit, we simply assert correlation. Again, as noted in the introduction, our features cannot identify songs with clever videos, terrific performers, or with a groundswell of social media support. But they can identify clever lyrics, and this alone does seem to be influential in the success of hit songs.

5.8 Conclusion: Lyrics Complexity and Craftmanship

We introduced our hit detection model in this chapter. We have used 31 rhyme, syllable and meter features for hit song detection, an important and largely unsolved, music information retrieval task. Our lyrics features significantly outperform 14 audio features for this task. Combing the lyrics and audio features gives us slightly better results.

We see that the presence of lots of rhymes, in particular complicated ones, makes it more likely that the song will be a hit. We assert correlation between the presence of these features and the probability of a song being a hit . The rhyme and meter features we use is indicative of craftmanship and the amount of effort put into songmaking. It is difficult to come up with audio features which can act as a proxy for the effort put in a song, and hence we believe that lyrics features are more powerful than the audio ones in discerning hits from flops. We argue that complex rhyme and meter is a detectable property of lyrics that indicates quality songmaking and artisanship and allows artists to become successful. An obvious drawback of this approach is that we cannot predict if the outcome of a cover or remake of a song is going to be any different from the original song, since they share lyrics.

Chapter 6

Lyrics Complexity and Shazam

6.1 Introduction

Shazam (shazam.com), is a music recognition service that analyzes a captured 10 second clip of a song, matching its acoustic fingerprint in a database of more than 11 million songs. It has been used to identify over 15 billion songs by its over 500 million users [23]. People shazam songs they like but do not know yet. Typically, these are songs someone else plays for them and for which they do not have access to the title and artist information. Shazam is presumably not used by listeners of streaming music services like Spotify or video services like YouTube where the artist or title information is needed if we want to listen to a song. It is unlikely that people listening to songs on their phones or media players would be shazaming a song since they just need to look at the screen to get the song information. Likely scenarios where people shazam songs could include songs played at a nightclub or gym or on the radio [58]. For a song to be shazamed it needs some exposure, and the number of shazams can be used to identify if an exposed song is being received favourably [58].

Why do people use a service like Shazam? Presumably, people shazam songs they like, but are unaware of, so that they can listen to them later. Such songs could be by relatively unknown upcoming artists. Popular artists have a distinct voice and style which the masses are already aware of. For example, “Let Her Go” by Passenger, who was relatively unknown before this song, has over 16 million shazams, while “Shake it Off” by Taylor Swift, one of the most successful current popular singers in the world, has just more than 3.5 million Shazams. Both these songs occupy roughly the same position on the 2014 Billboard Year-End Hot 100 chart [11].

In fact, this pattern is broader: on manual inspection, we see that a majority of the songs in the Shazam Hall of Fame [67] are first hits, where “first hit” of an artist is that artist’s first song to reach the Billboard’s Year-End Hot 100 singles chart. Also surprising is that many songs with lots of shazams do not in fact become hits at all (again defining hits as songs on the Billboard’s Year-End Hot 100 singles chart). For example, “Take Me to Church,” by Hozier and “Jubel,” by Klingande have never made it to the Billboard’s Year-End Hot 100 chart, though the former looks likely to do so for 2015.

Can we identify features that cause users to more likely to shazam a song? We solely focus on song lyrics in our answer, for all of the reasons discussed in our previous chapters. Additionally, songs with catchy lyrics may be more likely to rise above the background and have listeners shazam them. We use the complete set of 24 rhyme and syllable features of the Rhyme Analyzer [42], and add the seven new meter features identifying the fraction of lines written in a particular meter, as discussed in Chapter 5.

We hypothesize that we can separate hits from the Hall of Fame songs using these features. If so, perhaps the features separating the two classes could be used by relatively unknown artists to break out and become successful. These features make these more-often shazamed songs stand out from the background, and are part of the reason why people shazam them. Obviously, songs will still require a fair amount of exposure.

We use standard machine-learning algorithms, such as weighted support vector machine [19] and Bayesian network classifiers from Weka [35] to separate Shazam Hall of Fame and hit songs. We employ ten-fold cross validation. The SVM outperforms the Bayesian network, and is able to identify 54.41% of Hall of Fame songs and misclassify 18.51% hits as belonging to Hall of Fame. Our Bayesian Network suggests that lyrics with lots of rhymes, in particular complicated ones, are more likely to end up in the Shazam Hall of Fame. Using linear regression to build a model that can predict the number of shazams for a song, we find that models that include a measure of lyrical complexity that we define here as one of their features are better at predicting the number of shazams than models which does not use the lyrics complexity feature. Hence, this shows that complex lyrics does play an important role in determining whether a song is going to stand out and get noticed.

In Chapter 5, the focus is on craftsmanship and how writing complex lyrics with lots of rhymes, in particular complicated internal and imperfect rhymes, make it more likely that the song will end up becoming a hit. In this chapter, our focus is on noticeability or catchiness. We focus on how an upcoming artist can rise about the “sonic wallpaper,” get noticed and shazamed and become successful. Obviously, the common theme across both of these is that lyrics matter and complex rhyme and meter is a detectable property

of lyrics that indicates quality songmaking and artisanship and allows artists to become successful.

A limitation of focusing on lyrics is that it prevents us from distinguishing good and bad covers and remixes of the same song by different artists. Similarly, a song might be shazamed more because of catchy tune or beats but our method cannot detect that.

We are not aware of work on what makes a song catchy in the sense that Shazam capitalizes upon. An unique aspect of our work here is the advice we have for upcoming singers, regarding the rhyme usage in their lyrics they need to break out and become successful, which we present in Section 6.7.

6.2 Related Work

As mentioned in Chapter 5, Hirjee and Brown [40] came up with a probabilistic model to identify rhymes in song lyrics based on the different rhyming patterns found in hip-hop. In our work, we use the features of the Rhyme Analyzer [42] software from their work. Smith et al. [80], make use of TF-IDF weighting to find typical phrases and rhyme pairs in song lyrics and conclude that typical number one hits, on average, are more clichéd.

We in Chapter 5, have used the rhyme, meter and syllable features from the Rhyme Analyzer to separate hits from flops. We conclude that for purposes of hit identification, the rhyme complexity may in fact be a proxy for high-effort song writing and quality of artisanship

Matt Bailey, from Coleman Insights, [58] analyzed 16 weeks of U.S. Shazam charts and concluded that songs tend to peak on Shazam after sales but before radio exposure and on-demand listening. He concluded that the most high impact use of Shazam is to see which of the new exposed songs are sparking the greatest interest from listeners. He was inconclusive about Shazam’s ability to make a hit.

6.3 Data

As in Chapter 5, we define hits as songs from the Billboard Year-End Hot 100 singles chart in the years 2008-2014 not present in the Shazam Hall of Fame. The Shazam Hall of Fame has songs which have been shazamed over 5 million times. It is further divided into three categories: platinum (15 million+ shazams), gold (10 million+ shazams) and

silver (5 million+ shazams). At the time of the study there were 81 English-language and 1 Spanish-language song in the Hall of Fame. As noted earlier, most songs in the Hall of Fame are first hits by their artists such as Adele’s first hit, “Rolling in the Deep”. Some of the most popular current artists like Lady Gaga and Coldplay have no songs in the Hall of Fame.

We started with 81 English Hall of Fame songs and 508 hits. For all songs, we obtained lyrics from a free online lyrics repository (metrolyrics.com). We removed songs with lyrics less than 30 lines long since it is hard to predict rhyme features on such short lyrics, partly because they are also usually messy in their transcriptions [79]. We were left with 68 Hall of Fame songs and 389 hits. We downloaded Billboard Year-end charts from billboard.com while the number of shazams for a song was provided by Shazam.

6.4 Method

As described in Chapter 5, we use 31 rhyme, syllable and meter features. We used the complete set of 24 rhyme and syllable features of the Rhyme Analyzer [42] and added seven new meter features. These features separate lyrics from prose and could be considered proxy for skilled lyric writing, particularly given their previous use in separating hits from flops [79]. We refer the reader to Section 5.5 for more details on the features used.

We used a weighted-cost SVM, from LIBSVM [19], from Weka [35], with ten-fold cross validation. It assigns different misclassification cost to instances depending on the class they belong to which helps with unbalanced data sets. Tuning the misclassification costs, we can adjust the number of true and false positives: large and small values of misclassification cost give trivial classifiers, while intermediate costs trade false negatives for false positives. We also used the Bayesian network module from Weka with ten-fold cross validation, and we report the confusion matrices for the network that maximizes data likelihood.

6.4.1 Lyrics Complexity

We define lyrics complexity to be the sum of the normalized values of our rhyme, meter and syllable features. For each feature, its contribution to the total lyrical complexity measure ranges between 0 (if the feature is at its minimum value) and +1 (if the feature is at its maximum). We here have employed feature scaling, where each feature has the same weight to calculate the complexity of the song. Since we have 31 rhyme, meter and syllable features, the lyrical complexity of a song can range from 0 - 31. Using normalizing

scaling and considering only the features that matter in separating hits from the Hall of Fame songs (which we obtain from our Bayesian Network), we get very similar results, though the effect does seem stronger using the scaling described here.

Figure 6.2 and 6.3 shows the scatter plot showing the relationship between the number of shazams and lyrics complexity for the songs in the Billboard Year-End Hot 100 singles chart in the year 2013 and 2008 respectively. The 2013 scatter plot confirms our hypothesis: high-complexity songs also have high number of shazams, much better than the scatter plot obtained using the 2008 data.

6.5 Shazam Users

A common theme across our experiments is that our results improve in the years after 2011. We believe that this is because of the sudden jump in worldwide smartphone sales which led to mobile apps like Shazam becoming more ubiquitous and people shazaming more songs. This increase in shazaming has led to more and better quality data which in turn gives us better results. Table 6.1 shows the number of Shazam users and the number of smartphones sold from 2008 to 2014: from 2011 on, we see that the number of Shazam users is roughly 40% of sold smartphones.

Year	# of Shazam users	# of Smartphones sold
2008	20 million	139 million
2009	50 million	172 million
2010	100 million	296 million
2011	165 million	472 million
2012	250 million	680 million
2013	300 million	969 million
2014	500 million	1244 million

Table 6.1: The number of Shazam users and the number of smartphones sold worldwide.

6.6 Results

We attempt to separate Hall of Fame songs from hits using rhyme, syllable and metre features using a Bayesian Network and an SVM. Using linear regression models we show

that lyrical complexity is a vital feature for predicting the number of times a song will be shazamed. Furthermore, we show that Hall of Fame songs are lyrically much more complicated than hits not found in the Hall of Fame.

Lyrics features do a good job of separating hits from Hall of Fame songs. We get the best results using a weighted-cost SVM, where we are correctly able to identify 54.41% of Hall of Fame songs and misclassify 18.51% hits as belonging to Hall of Fame. Section 6.6.1 details these experiments.

Using linear regression to build a model which can predict the number of shazams for a song we find that models which include lyrics complexity as one of its features has a lower Akaike information criterion (AIC) score than the model which does not use the lyrics complexity feature. These results are found in Section 6.6.2. The improvement of results in the years after 2011 is a common theme across all our experiments, which again we expect is a consequence of Shazam’s large user base that we described in Section 6.5.

We observe that songs by upcoming artists usually have chorus as the lyrically least complex part of the song while songs by established artists usually have chorus as the lyrically most complex part of the song. We discuss these results in Section 6.6.3.

	Weighted-cost SVM
# correctly classified Hall of Fame	37
# misclassified Hall of Fame	31
# correctly classified Hits	317
# misclassified Hits	72
precision(Hall of Fame)	0.339
recall(Hall of Fame)	0.544
F-score(Hall of Fame)	0.418

Table 6.2: The results for separating Shazam Hall of Fame songs from hits using a weighted-cost SVM

6.6.1 Separating the Shazam Hall of Fame songs from Hits using the lyrics features

In our first experiment, we separate Hall of Fame songs from hits using rhyme, syllable and meter features and standard machine learning algorithms (Bayesian Networks and SVMs). The hypothesis is that Shazam Hall of Fame songs are lyrically more complex than hits. The results obtained using a weighted-cost SVM and a Bayesian Network are shown in

Tables 6.2 and 6.3 respectively. A weighted-cost SVM outperforms Bayesian networks in detecting Hall of Fame songs with better precision and recall values. As in Chapter 5, we assign different penalties to misclassified instances of different classes. A weight of 8 implies that it is 8 times more expensive to misclassify a Hall of Fame song (a false negative) than a hit (a false positive) from the training set. We obtain the best result when using a weight of 3.4, for which results are in Table 5.6: the weight parameter is tuned to see what the precision is when the recall is close to 50%.

	Bayesian Network
# correctly classified Hall of Fame	19
# misclassified Hall of Fame	49
# correctly classified Hits	356
# misclassified Hits	33
precision(Hall of Fame)	0.365
recall(Hall of Fame)	0.279
F-score(Hall of Fame)	0.316

Table 6.3: The results for separating Shazam Hall of Fame songs from hits using a Bayesian Network. The SVM does a better job at predicting hits, though the Bayes Net classifier is more stringent

Shazam Hall of Fame songs are lyrically more complicated, though other features do play some role. From our Bayesian network we find that our most important features are features like syllables per line, end pairs shrunk and syllable variation which highlight line length and its variation across the lyrics; features like rhymes per line and rhymes per syllable which account for the frequency of rhymes in the lyrics; and features like link rhymes per line, bridge rhymes per line and compound rhymes per line highlight the complex rhyming strategy. As in Chapter 5, the Bayesian network we obtain from Weka is naïve Bayes: the effect of one feature is independent of another. The definition of these features is in Table 5.4.

In our earlier work, presented as Chapter 5 of this thesis, we concluded that hits are lyrically more complicated than flops, having frequent complicated rhymes. Hall of Fame songs are even more complicated than hits. Since most of the songs in the Hall of Fame are first hits, we conclude that for an upcoming artist to break out and become successful, one good route is to write lyrically complicated lyrics with lots of complex rhymes.

There are some first hits with high chart position but surprisingly few shazams, such as “Whatever it is,” by Zac Brown Band. It is his first hit, occupies the 94th position on

the 2008 Billboard Year-End Hot 100 singles chart but has less than 500,000 shazams, a small number for a hit song. This could possibly be because of a lack of exposure, which is a prerequisite for a song to be shazamed.

Another interesting case is that of “Troublemaker” Olly Murs’s first hit occupies the 82nd position on the 2013 Billboard Year-End Hot 100 singles chart and has over 3.5 million shazams, a good number for a song on the lower half of the chart. The song features Flo Rida, an established artist whose distinct voice and style is recognized by the masses. We believe that this song did not make it to the Hall of Fame because of Flo Rida’s presence.

6.6.2 Lyrics complexity and shazams

We define lyrics complexity to be the sum of the normalized values of our rhyme, meter and syllable features, where all features are scaled to between 0 (the minimum value for the feature) and 1 (the maximum value). In this metric, lyrics with more rhymes, especially complicated ones, will have higher scores. We were curious to determine if lyrics complexity was a relevant feature in a linear regression model that attempted to predict the number of shazams. The considerable growth in Shazam’s user base along with rising smart phone penetration makes the year of release of the song a vital feature of any model. As argued earlier, people tend to shazam songs by relatively unknown artists more than established ones, hence it is highly likely that an artist’s first few major hits will have more shazams than his other songs. Hence, whether a given song was released in the same year as an artist’s first hit is another variable in our model. A popular song will be shazamed more than a flop, hence our final feature is the position of the song in the Billboard’s Year-End Hot 100 chart. Our first model uses these three features: year, rank and whether a song was released in the year of first hit. The second model has an additional lyrics complexity feature in addition to the features of the previous model. If the model with lyrics complexity is better at predicting shazams, then we can conclude that the lyrical complexity does play an important role in determining the number of shazams, or is correlated with another important feature not captured by our other features.

The Akaike information criterion (AIC) [1] is a measure of the relative quality of a statistical model for a given set of data. It rewards goodness of fit, but it also includes a penalty that is an increasing function of the number of estimated parameters to discourage overfitting. The preferred model is the one with the minimum AIC value. The model with the lyrics complexity feature has a lower AIC value (7956.96 vs 7987.99) and hence adding the lyrics complexity feature gives a better model. There is a positive correlation between the features year of release of song, lyrics complexity, the year of first hit by artist

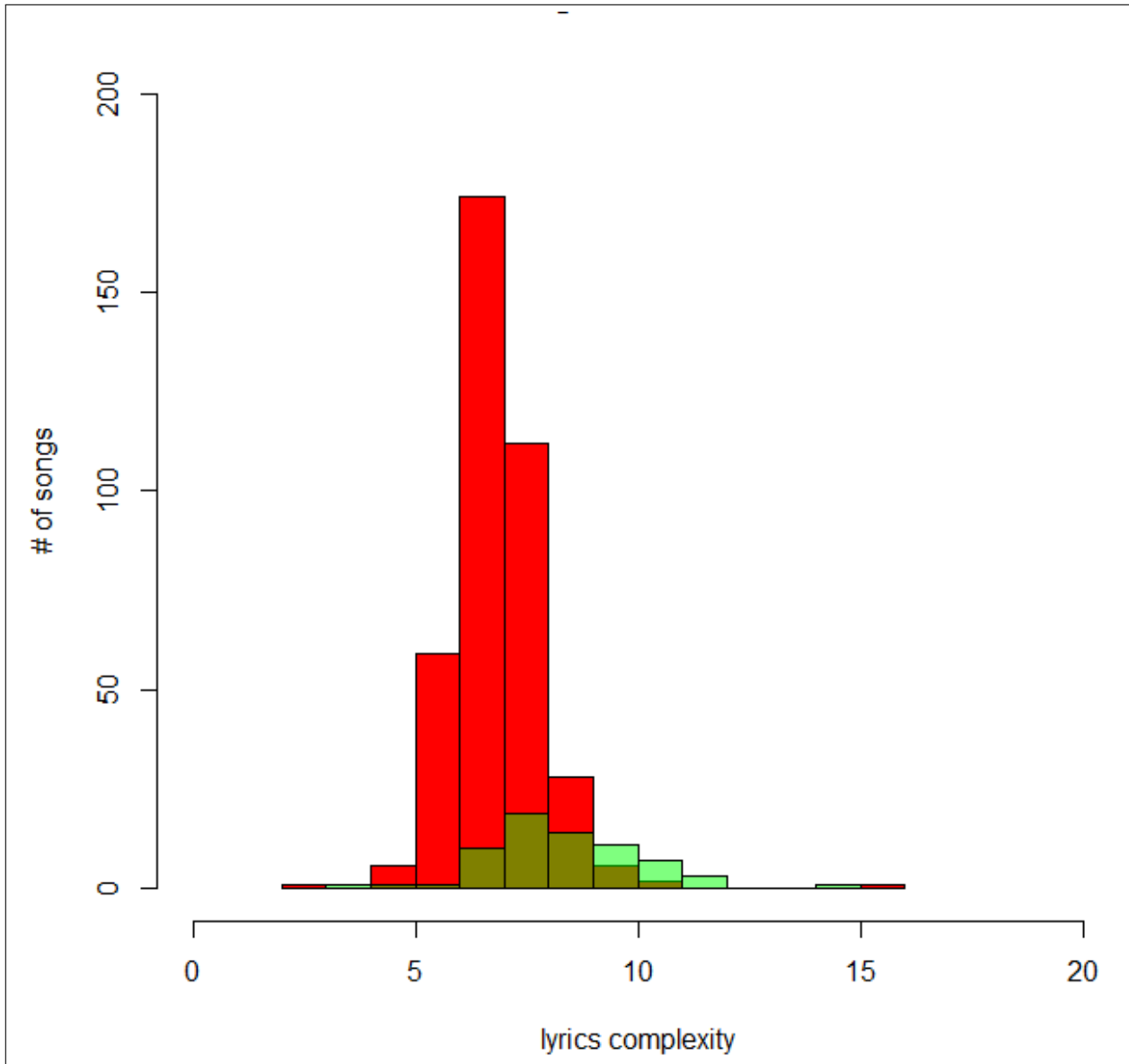


Figure 6.1: An overlapped histogram of the lyrics complexity of the hits and Hall of Fame songs. The red coloured part is the histogram for the Hall of Fame songs while the histogram for hits is coloured green. Despite having far more hits than Hall of Fame songs, the majority of songs with lyric complexity measure greater than 8 are from the Hall of Fame. The overlapped portion is coloured dark green.

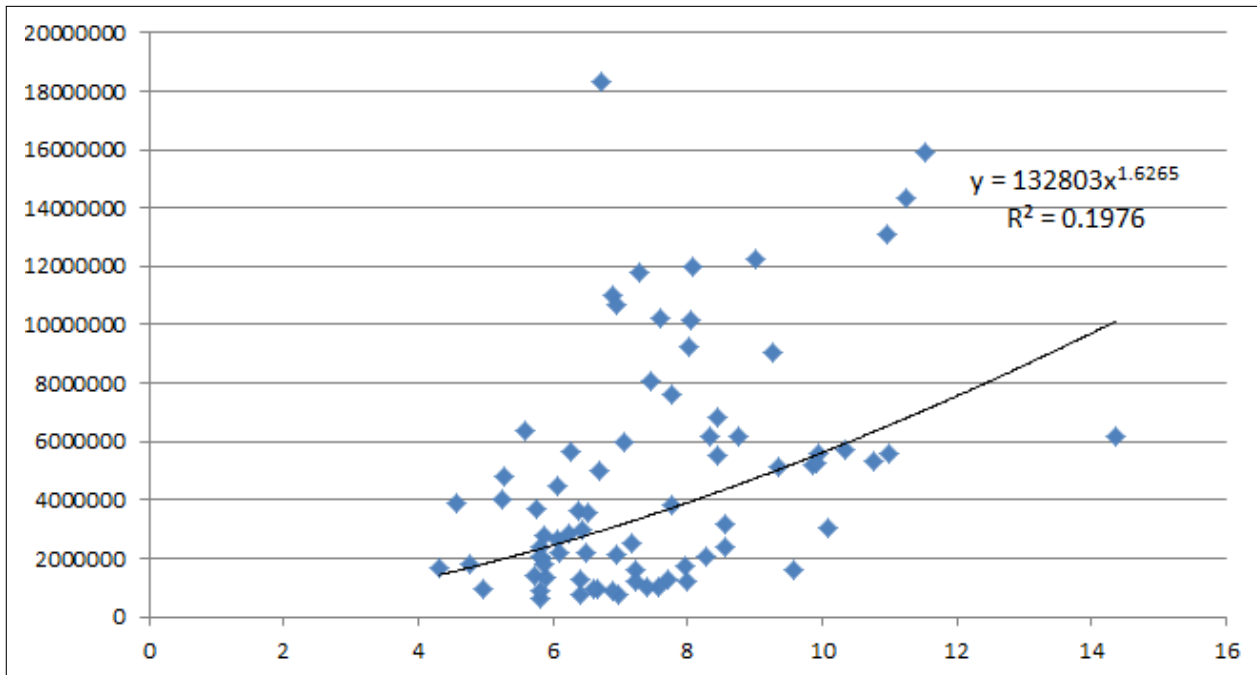


Figure 6.2: A scatter plot showing the relationship between the number of shazams and lyrics complexity for the songs in the 2013 Billboard Year-End Hot 100 singles chart (Correlation coefficient = 0.3986). The curve of best fit has R-squared value of 0.1976. Lyrically complex songs tends to be shazamed more.

and the number of shazams. Lyrically-complicated songs with lots of complex rhymes will more likely be shazamed more than simpler songs with fewer rhymes. Popular songs by relatively unknown artists will be shazamed more than songs by established and famous artists.

Figure 6.1 shows an overlapping histogram for the hits and the Hall of Fame songs in our data set. From the figure it is clear that the Hall of Fame songs are lyrically more complicated than the hits, where the complexity of a song is the sum of its normalized feature values. The mean complexity of hits in our data set is 6.84 while a Shazam Hall of Fame songs have a mean complexity of 8.35. In Chapter 5, we showed that hits are lyrically more complicated than flops; here we show that Shazam Hall of Fame songs are even more complicated than hits. We can further confirm this positive correlation between shazams and lyrics complexity via Figure 6.2, which is the scatter plot between the number of shazams and the lyrics complexity for the songs which made it to the 2013 Billboard Year-End Hot 100 singles chart. High-complexity songs also have relatively high number of

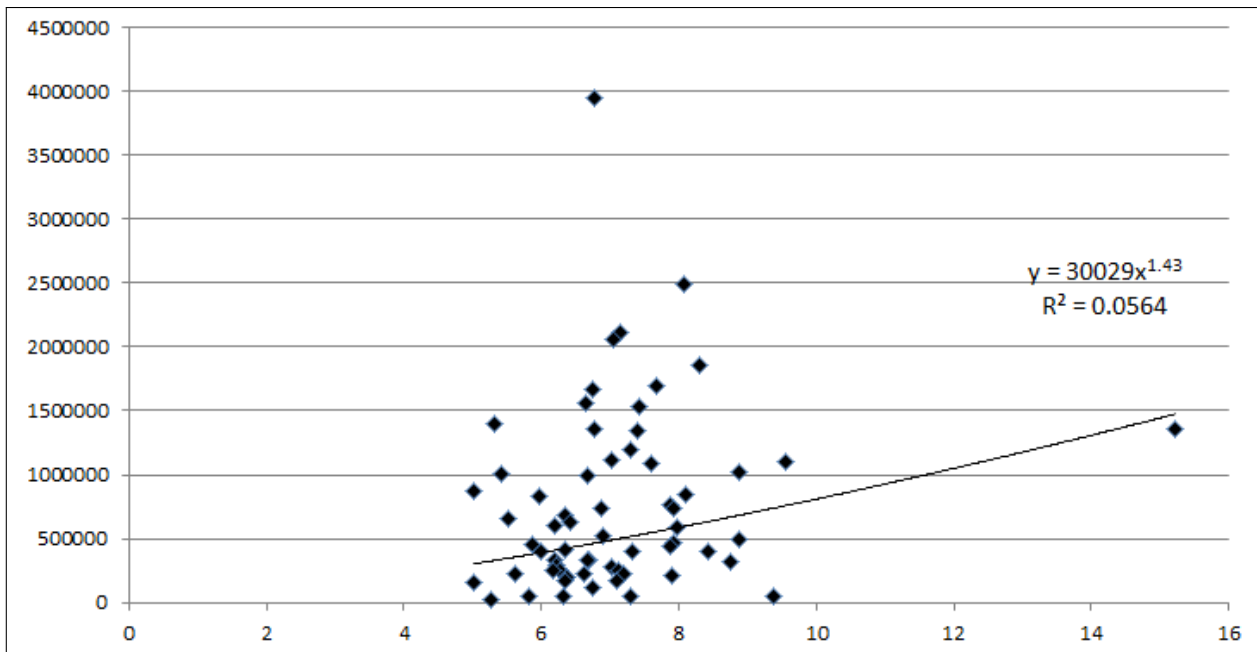


Figure 6.3: A scatter plot showing the relationship between the number of shazams and lyrics complexity for the songs in the 2008 Billboard Year-End Hot 100 singles chart (Correlation coefficient = -0.1757). The curve of best fit has R-squared value of 0.0564. In the early years, the relationship between shazams and lyric complexity is less clear.

shazams. Table 6.4 lists the top five most lyrically-complicated songs in the Hall of Fame, the majority of which belong to the platinum and gold category of the Hall of Fame. Some exceptions are lyrically complex songs belonging to famous artists; as discussed earlier, these songs are not shazamed as often as songs by relatively unknown artists. Examples include, “Don’t Stop The Music”, a lyrically-complex song by Rihanna, has barely over a million shazams, presumably because of Rihanna’s popularity and distinctive voice.

Song	Artist	Year	Complexity	Shazams
Some Nights	Fun	2012	14.35	6,205,894
Let Her Go	Passenger	2012	11.52	16,013,361
Thrift Shop	Macklemore & Ryan Lewis	2012	11.24	14,397,537
Clarity	Zedd	2012	11.00	5,602,391
Can’t Hold Us	Macklemore & Ryan Lewis	2012	10.96	13,206,552

Table 6.4: The top five most lyrically-complicated Hall of Fame songs.

Figure 6.4 shows the relationship between the number of shazams and the 2014 Billboard Year-End Hot 100 rank. Again, the general trend in the graph is that a bigger hit is shazamed more, but the outliers are the interesting songs. These songs are usually the first hit of an upcoming artist and hence the spike in interest. Examples of this phenomenon are “Let Her Go” by Passenger, the 19th position in the 2014 Billboard Year-End Hot 100 chart, but is one of the most shazamed songs of all time. Similarly, “La La La” by Naughty Boy occupies the 82nd spot on the chart but has close to 9 million shazams. “Blurred Lines” was the #2 song in 2013, which explains its outlier status for 2014. Songs by established artists are usually shazamed in proportion to their rank on the chart, “Maps” by Maroon 5, an established artist, has close to 4 million shazams and occupies the 19th place in the chart. Similarly, “Dark Horse” by Katy Perry occupies the 2nd position and has over 10 million shazams.

6.6.3 Shazam Hall of Fame and Chorus Complexity

A majority of the songs by upcoming artists in the Shazam Hall of Fame have songs where the chorus is the least lyrically complicated part of the song. Examples include “Take Me to Church,” by Hozier and “All Of me,” by John Legend. Of course, there are exceptions like “Let Her Go,” by Passenger, where the chorus is the most lyrically complicated part of the song. On the other hand, songs by established singers in the Hall of Fame usually have the chorus as the most lyrically complex part of the song. Examples include “Counting

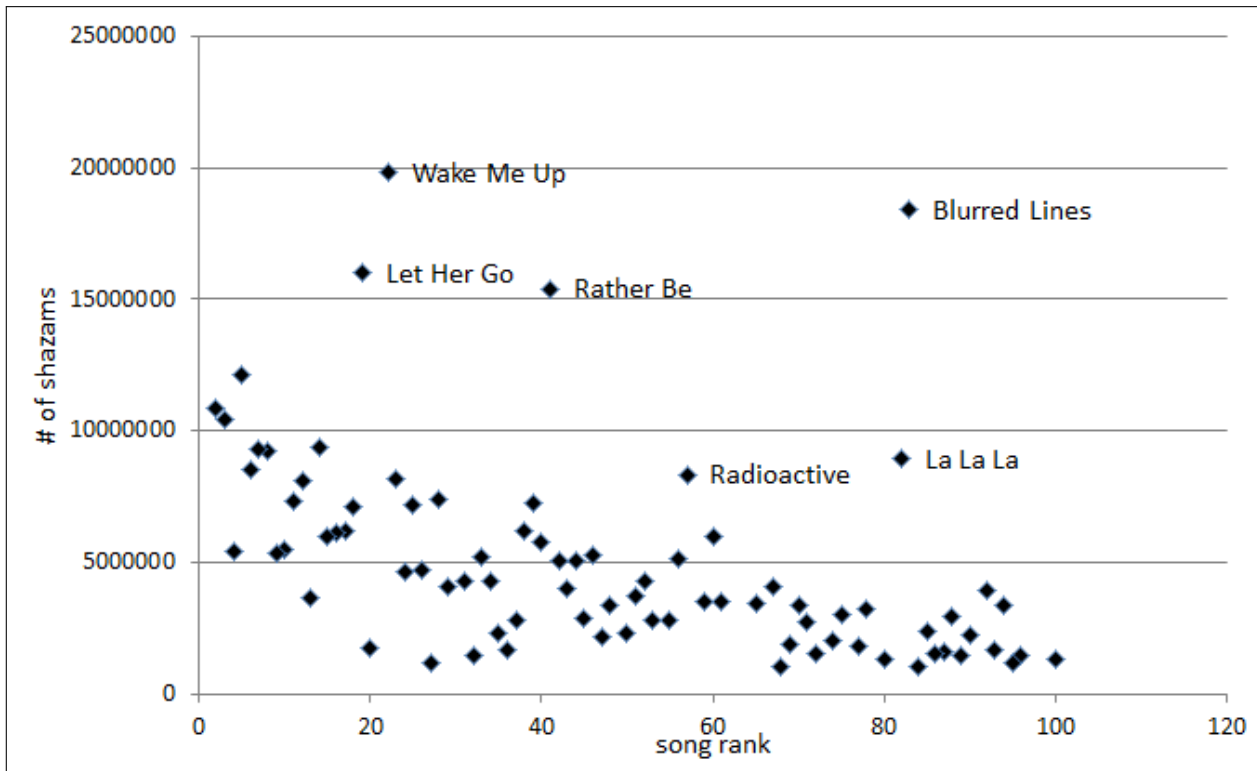


Figure 6.4: A scatter plot showing the relationship between the number of shazams and the 2014 Billboard Year-End Hot 100 rank for the songs which made it to the year end chart. “Blurred Lines” was the #2 song in 2013, which explains its outlier status for 2014.

Stars” by One Republic and “Demons” by Imagine Dragons. Perhaps this is a common pattern imposed (presumably unconsciously) when an artist becomes famous and signs up for big labels. Table 6.5 lists Hall of Fame songs having the chorus as the least and most lyrically complicated part of the song.

Complicated chorus	Less Complex chorus
Thrift Shop	One More Night
Take Me to Church	When I Was Your Man
Can’t Hold Us	Counting Stars
All of Me	Just Give Me a Reason
Stereo Love	Good Feeling

Table 6.5: The table lists Hall of Fame songs having the chorus as the lyrically most complex and least complex part of the song.

6.7 Conclusion: Advice for musicians

Traditionally, rhyme has always been believed to be an important part of lyrics, which separates it from prose. Some lyricists like Stephen Sondheim prefer perfect rhymes, where vowels and the succeeding consonant are similar, over near rhymes, where either the vowel sound or the consonant sound differ [81]. They believe that perfect rhymes are better than near rhymes and are indicative of the efforts put in by the lyricist [81]. Sondheim acknowledges that other lyricists are of the view that true rhymes allow a limited range to express ones feelings, giving the inclination to use near rhymes [81].

Present day pop and rock music are full of near rhymes [73], which some believe to be a decline in music quality [81]. It is possibly because of the belief that an artist’s best work is contained in his or her first hit album, after which he usually reverts back to writing lyrically less complicated songs [7]. This could be because of the pressure by big labels who might want him to produce a certain kind of music [17].

An example of a successful artist coming up through the route of highly-shazamed lyrically complex songs is Macklemore. He is not associated with any music label, and three of his four hits are in the Shazam Hall of Fame. They are more complicated than the average hit. “White Walls”, his only hit not in the Hall of Fame occupies the 92nd place in the 2014 Billboard’s Year-End singles chart but has close to 4 million shazams, an extremely high number for a low-ranked song. We note that Macklemore is a controversial

figure in the rap press where some believe him not to be a true rapper [6]. Nonetheless, he is famous and successful and has a massive fan following [18].

Songs in the Hall of Fame are lyrically more complicated than hits and as argued earlier, a majority of them are by upcoming artists. Hence, this suggests one way for an upcoming artist to succeed is to write complicated lyrics with lots of rhymes, especially complicated ones, to get recognized, and rise above the sonic wallpaper of daily life, be shazamed, and become successful. Our focus, in this study, has been on imperfect and internal rhymes and their presence does improve a songs chance of making it to the Hall of Fame.

Chapter 7

Conclusion

In this thesis, we have studied the usefulness of lyrics, which compared to the audio, have largely been ignored by the Music Information Retrieval (MIR) research [63], in solving some important MIR tasks. We find that lyrics do matter and in some cases outperform audio features in solving MIR tasks. We find that the presence of lyrics has a variety of significant effects on how people perceive songs. We have also shown that we can separate lyrics and poetry based on the kinds of adjectives used. Our most surprising result is that lyrically complicated songs with lots of rhymes, in particular internal and imperfect rhymes, are more likely to end up as hits. We see that the songs in the Shazam Hall of Fame are lyrically even more complicated than these hits.

In Chapter 3 we presented the results from a user study to determine if lyrics can help bridge the music mood perception between different cultures. Our experiment shows that the presence of lyrics has a significant effect on how people perceive songs. To our surprise, reading lyrics alongside listening to a song does not significantly reduce the differences in music mood perception between Canadian and Chinese listeners. Also, while we included two different sets of Canadian listeners (Canadian-Chinese, and Canadians not of Chinese origin), we can make no useful conclusions about the Chinese-Canadian group. We do consistently see that presence of both audio and lyrics reduces the consistency of music mood judgment between Chinese and Canadian listeners. This phenomenon may be because of irony caused by negative words presented in proximity to upbeat beats, or it could be that presenting both audio and lyrics together might be a completely different experience for the listener. This is an obvious setting for further work. We have shown that the mood of a song depends on its experiential context. Interestingly, songs where listeners agree strongly about the mood of the song when only listening to the recording are often quite uncertain in their mood assignments when the lyrics are shown alongside the recording.

We also show that many “melancholy” lyrics are found in songs assigned to a more cheerful mood by listeners, again suggesting that for such songs, the extent to which listeners focus on the lyrics may influence how sad they view a song to be. We analyzed the mood assignments of participants on rock and hip-hop songs. We see that people tend to agree much more to the mood of a hip-hop song when they are made to listen to the song. We found that for rebellious or negative rock songs, being able to read lyrics leads to more agreement in music mood but being able to hear the audio leads to more agreement for positive songs. In both the genres we found that hearing audio while reading lyrics lead to less agreement on music mood of songs. Our results suggest that music mood is so dependent on cultural and experiential context to make it difficult to claim it as a true concept. With the classification accuracy of mood classification systems reaching a plateau with no significant improvements we suggest that we need to redefine the term “music mood” and change our approach toward the music mood classification problem. We fundamentally also wonder if “mood” as an MIR concept needs to be reconsidered. If listeners disagree more or less about the mood of a song when it is presented alongside its lyrics, that suggests a general uncertainty in the concept of “mood”. We leave more evidence gathering about this concept to future work as well.

A possible extension to our work could be running a similar study using a larger set of songs and more participants, possibly from more diverse cultures than the ones we studied. Future studies could focus on multi-modal music mood classification where a song could belong to more than one mood, to see if even in this more robust domain there is a stable way to assign songs to clusters of moods when they are experienced in different contexts. We also wonder if other contextual experiments can show other effects about mood: for example, if hearing music while in a car or on public transit, or in stores, makes the “mood” of a song more uncertain.

In Chapter 4, we developed a method to detect the genre of a document based on the probability distribution of synonymous adjectives. Our key finding is that the choice of synonym for even a small number of adjectives are sufficient to reliably identify the genre of documents. In accordance with our hypothesis, we show that there exist differences in the kind of adjectives used in different genres of writing. We calculate the probability distribution of synonymous adjectives over the three kinds of documents and using this distribution and a simple algorithm, we are able to distinguish among lyrics, poetry and article with an accuracy of 67%, 57% and 80% respectively. Adjectives likely to be used in lyrics are more rhymable than the ones used in poetry. This might be because lyrics are written keeping in mind the melody, rhythm, instrumentation, quality of the singer’s voice and other qualities of the recording while poetry is without such concerns. There is no significant difference in the semantic orientation of adjectives which are more likely to be

used in lyrics and those which are more likely to be used in poetry. Using the probability distributions, obtained from training data, we present adjectives more likely to be used in lyrics rather than poetry and vice versa for twenty common concepts. Using the probability distributions and our algorithm we show that we can discern poetic lyricists like Bob Dylan and Stephen Sondheim from nonpoetic ones like Bryan Adams and Kesha. Our algorithm consistently misclassifies a majority of the lyrics of such poetic lyricists as poetry while the percentage of misclassified lyrics as poetry for the non-poetic lyricists is significantly lower.

Calculating the probability distribution of adjectives over the various document types is a vital step in our method which in turn depends on the synonyms extracted for an adjective. Synonym extraction is still an open problem and with improvements in it our algorithm will give better accuracy levels. We extract synonyms from three different sources: Wikipedia, WordNet and an online thesaurus, and prune the results based on the semantic similarity between the adjectives and the obtained synonyms. We use a simple naïve algorithm, which gives us a better result than Naïve Bayes. An extension to the work can be coming up with an improved version of the algorithm with better accuracy levels. Future works can use a larger dataset for lyrics and poetry (we have an enormous dataset for articles) to come up with better probability distributions for the two document types or to identify parts of speech that effectively separates genres of writing. Our work here can be extended to different genres of writings like prose, fiction etc. to analyze the adjective usage in those writings. It would be interesting to do similar work for verbs and discern if different words, representing the same action, are used in different genres of writings.

In Chapter 5, we introduce our hit detection model. We have used 31 rhyme, syllable and meter features for hit song detection, an important music information retrieval task. Our lyrics features significantly outperform 14 audio features for this task. Combing the lyrics and audio features gives us slightly better results. We select hits to be songs which made it to the Billboard Year-End Hot 100 singles between the years 2008 and 2013. Flops are non-hit songs, depending on our definition of flop, ranging from a very broad one to extremely restricted ones. Our largest data set consists of 492 hits and 6323 flops by the most popular current English-language music artists. We use Bayesian networks and weighted-cost support vector machines with 10-fold cross validation. Varying the weights of the SVM, we can adjust the values of true and false positives depending on the economic costs associated with missing a hit and investing in a flop. For our largest data set, using just the lyrics features we can identify about half of the hits, while misclassifying only 12.8% of flops as hits. For the hit detection task, we are consistently able to correctly identify about half of the hits across all the four definitions of flops.

We see that the presence of many rhymes, in particular complicated ones, makes it

more likely that the song will be a hit. Surprisingly, very loud songs are more likely to be flops. We do not claim that the presence of these features make a hit, though we do assert correlation. The rhyme and meter features we use is indicative of craftsmanship and the amount of effort put into songmaking. It is difficult to come up with audio features which can act as a proxy for the effort put in a song, and hence we believe that lyrics features are more powerful than the audio ones in discerning hits from flops. An obvious drawback of this approach is that we cannot predict if the outcome of a cover or remake of a song is going to be any different from the original song, since they share lyrics. Our work is novel and simple, and it outperforms previous hit detection models. An extension might be to combine these features with features derived from either recordings, scores or text complexity, and to focus on specific genres.

In Chapter 6, we discuss a way for upcoming artist to rise above the “sonic wallpaper” and become successful through the route of highly-shazamed lyrically complex songs. Shazam is used by people to identify songs they like, but are unaware of, usually in scenarios where someone else plays the song for them. Such songs are usually by relatively unknown and upcoming artists as popular artists have distinct voice and style which the masses are usually aware of. We wanted to identify whether lyric features could identify songs likely to be heavily shazamed, as complex rhyme and meter might make a song surprising enough that a person would use the app on it. We select hits to be songs which made it to the Billboard Year-End Hot 100 singles between the years 2008 and 2014 and are not present in the Shazam Hall of Fame. Our data set consists of 68 Hall of Fame songs and 389 hits. We use a Bayesian Network and a weighted-cost support vector machine to separate hits from Hall of Fame songs using the 31 rhyme, meter and syllable features as we used in Chapter 5. Consistent with our hypothesis, we see that the presence of lots of rhymes, in particular complicated imperfect and internal rhymes, makes it more likely that the song will end up in the Shazam Hall of Fame. We define lyrics complexity to be the sum of the normalized values of our 31 lyrics features. Lyrics complexity is an important feature of a linear regression model used to predict the number of shazams. Outliers on the shazams versus Billboard Year-End Hot 100 rank graph are usually hits by upcoming artists. Songs by upcoming artists usually have the chorus as the lyrically least complex part of the song while songs by established artists usually have the chorus as the lyrically most complex part of the song.

Our work is simple and is surprisingly effective at explaining which songs will be shazamed, and how often. Our advice to upcoming singers is to write lyrically complicated songs with many complicated rhymes in order to rise above the sonic wallpaper, get noticed and shazamed, and become famous.

In this thesis, we show the usefulness of lyrics in solving four important Music Informa-

tion Retrieval tasks. For our first result, we show that the presence of lyrics has a variety of significant effects on how people perceive songs. We then proceed to show that we can predict the genre of a document based on the adjective choices made by the authors. For our next result, we show that rhyme and meter features are useful in separating hits and flops and outperform the currently popularly used audio features. Using the same features we can also detect songs which are likely to be shazamed heavily. We argue that complex rhyme and meter is a detectable property of lyrics that indicates quality songmaking and artisanship and allows artists to become successful.

References

- [1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [2] Craig A Anderson, Nicholas L Carnagey, and Janie Eubanks. Exposure to violent media: the effects of songs with violent lyrics on aggressive thoughts and feelings. *Journal of Personality and Social Psychology*, 84(5):960–971, 2003.
- [3] Ion Androutsopoulos, John Koutsias, Konstantinos V Chandrinos, and Constantine D Spyropoulos. An experimental comparison of naïve Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–167. ACM, 2000.
- [4] Ani DiFranco Wikipeda article. http://en.wikipedia.org/wiki/Ani_DiFranco.
- [5] Beard, Alison. Lifes Work: Annie Lennox. <https://hbr.org/2010/10/lifes-work-annie-lennox>. 2010.
- [6] Ben Beaumont-Thomas. The Grammys 2014: is Macklemore’s success bad for hip-hop? *The Guardian*, 27 January 2014.
- [7] Ben Kaye. Proof the sophomore album slump is a real problem. <http://consequenceofsound.net/2015/02/proof-the-sophomore-album-slump-is-a-real-problem/>. 2015.
- [8] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *International Society for Music Information Retrieval*, pages 591–596, 2011.

- [9] Mireille Besson, Frederique Faita, Isabelle Peretz, A-M Bonnel, and Jean Requin. Singing in the brain: Independence of lyrics and tunes. *Psychological Science*, 9(6):494–498, 1998.
- [10] Billboard. <http://www.billboard.com/>.
- [11] Billboard Year-End Hot 100 singles of 2014. <http://www.billboard.com/charts/year-end/2014/hot-100-songs>.
- [12] Dean Leonard Biron. Writing and Music: Album Liner Notes. *PORTAL Journal of Multidisciplinary International Studies*, 8(1), 2011.
- [13] Kerstin Bischoff, Claudiu S Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. Social knowledge-driven music hit prediction. In *Proceedings of the Advanced Data Mining and Applications Conference*, pages 43–54. 2009.
- [14] Christian Bøhn and Kjetil Nørvåg. Extracting named entities and synonyms from Wikipedia. In *IEEE International Conference on Advanced Information Networking and Applications*, pages 1300–1307, 2010.
- [15] Bruce Springsteen Wikipedia article. http://en.wikipedia.org/wiki/Bruce_Springsteen.
- [16] Felix Budelmann. Introducing Greek lyric. *Budelmann, F.(red.). The Cambridge Companion to Greek lyric. Bladsy*, pages 1–18, 2009.
- [17] David Byrne. *How music works*. McSweeney’s, 2012.
- [18] Jon Caramanica. Finding a place in the hip-hop ecosystem. *The New York Times*, page C1, 27 January 2014.
- [19] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- [20] Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233, 2006.
- [21] Tina Chui, Kelly Tran, and John Flanders. Chinese Canadians: Enriching the cultural mosaic. *Canadian Social Trends*, 76(2), 2005.
- [22] Michael Deacon. Morrissey doesn’t write poetry, he writes lyrics. *The Telegraph*, 23 May 2009.

- [23] Dean Van Nguyen. Silicon Republic. <http://www.siliconrepublic.com/digital-life/item/38714-15-billion-songs-have-been/>. 2014.
- [24] Ruth Dhanaraj and Beth Logan. Automatic Prediction of Hit Songs. In *International Society for Music Information Retrieval*, pages 488–491, 2005.
- [25] Ani DiFranco. *Ani DiFranco: Verses*. Seven Stories Press, 2011.
- [26] J Downie, Kris West, Andreas Ehmman, and Emmanuel Vincent. The 2005 music information retrieval evaluation exchange (mirex) 2005): Preliminary overview. In *International Conference for Music Information Retrieval*, pages 320–323, 2005.
- [27] J Stephen Downie, Donald Byrd, and Tim Crawford. Ten years of ISMIR: Reflections on challenges and opportunities. In *International Society for Music Information Retrieval*, pages 13–18, 2009.
- [28] H Elovitz, Rodney Johnson, Astrid McHugh, and J Shore. Letter-to-sound rules for automatic translation of English text to phonetics. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 24, pages 446–459, 1976.
- [29] Doris R Entwisle and Catherine Garvey. Verbal productivity and adjective usage. *Language and Speech*, 15(3):288–298, 1972.
- [30] Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of Language Resources and Evaluation Conference*, volume 6, pages 417–422, 2006.
- [31] Jianyu Fan and Michael A Casey. Study of Chinese and UK hit Songs Prediction. *Proceedings of Computer Music Multidisciplinary Research (CMMR)*, 2013.
- [32] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [33] Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. Poetic statistical machine translation: rhyme and meter. In *Conference on Empirical Methods in Natural Language Processing*, pages 158–166, 2010.
- [34] Nicolas Guéguen, Céline Jacob, and Lubomir Lamy. love is in the air: Effects of songs with romantic lyrics on compliance with a courtship request. *Psychology of Music*, 38(3):303–307, 2010.

- [35] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. In *ACM SIGKDD Explorations Newsletter*, volume 11, pages 10–18, 2009.
- [36] Hui He, Jianming Jin, Yuhong Xiong, Bo Chen, Wu Sun, and Ling Zhao. Language feature mining for music emotion classification via supervised learning from lyrics. In *3rd International Symposium on Computation and Intelligence*, pages 426–435. Springer, 2008.
- [37] David H Henard and Christian L Rossetti. All you need is love? Communication insights from pop music’s number-one hits. *Journal of Advertising Research*, 54(2):178–191, 2014.
- [38] Dorien Herremans, David Martens, and Kenneth Sørensen. Dance hit song prediction. In *International Workshop on Machine Learning and Music, ECML/PKDD*, 2013.
- [39] Hussein Hirjee. Rhyme, rhythm, and rhubarb: Using probabilistic methods to analyze hip hop, poetry, and misheard lyrics. Master’s thesis, University of Waterloo, 2010.
- [40] Hussein Hirjee and Daniel G Brown. Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics. In *International Society for Music Information Retrieval*, pages 711–716, 2009.
- [41] Hussein Hirjee and Daniel G Brown. Automatic detection of internal and imperfect rhymes in rap lyrics. In *International Society for Music Information Retrieval*, pages 711–716, 2009.
- [42] Hussein Hirjee and Daniel G Brown. Rhyme Analyzer: An Analysis Tool for Rap Lyrics. In *International Society for Music Information Retrieval*, 2010.
- [43] Hussein Hirjee and Daniel G. Brown. Using automated rhyme detection to characterize rhyming style in rap music. *Empirical Musicology Review*, 5(4):121–145, 2010.
- [44] Xiao Hu and J Stephen Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *International Society for Music Information Retrieval*, pages 67–72, 2007.
- [45] Xiao Hu and J Stephen Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 159–168. ACM, 2010.

- [46] Xiao Hu and J Stephen Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *International Society for Music Information Retrieval*, pages 619–624, 2010.
- [47] Xiao Hu, J Stephen Downie, Cyril Laurier, Mert Bay, and Andreas F Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *International Society for Music Information Retrieval*, pages 462–467, 2008.
- [48] Xiao Hu and Jin Ha Lee. A cross-cultural study of music mood perception between American and Chinese listeners. In *International Society for Music Information Retrieval*, pages 535–540, 2012.
- [49] Yajie Hu, Xiaou Chen, and Deshun Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *International Society for Music Information Retrieval*, pages 123–128, 2009.
- [50] Poem Hunter. <http://www.poemhunter.com>.
- [51] Long Jiang and Ming Zhou. Generating Chinese couplets using a statistical MT approach. In *International Conference on Computational Linguistics*, volume 1, pages 377–384, 2008.
- [52] Katerina Kosta, Yading Song, György Fazekas, and Mark B Sandler. A study of cultural dependence of perceived mood in Greek music. In *International Society for Music Information Retrieval*, pages 317–322, 2013.
- [53] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. Multimodal music mood classification using audio and lyrics. In *International Conference on Machine Learning and Applications*, pages 688–693, 2008.
- [54] Jin Ha Lee, M Cameron Jones, and J Stephen Downie. An analysis of ISMIR proceedings: Patterns of authorship, topic, and citation. In *International Society for Music Information Retrieval*, pages 57–62, 2009.
- [55] Heidi I Brummert Lennings and Wayne A Warburton. The effect of auditory versus visual violent media exposure on aggressive behaviour: the role of song lyrics, video clips and musical tone. *Journal of Experimental Social Psychology*, 47(4):794–799, 2011.
- [56] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):5–18, 2006.

- [57] Metro Lyrics. <http://www.metrolyrics.com>.
- [58] Matt Bailey. What You Can and Can't Learn From Shazam. <http://www.colemaninsights.com/news/what-you-can-and-cant-learn-from-shazam>. 2015.
- [59] George A Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- [60] Pharoahe Monch. Internal affairs. *Rawkus Records*, 1999.
- [61] Keith Negus. *Bob Dylan*. Equinox London, 2008.
- [62] Echo Nest. <http://echonest.com/>.
- [63] Yizhao Ni, Raúl Santos-Rodríguez, Matt McVicar, and Tijl de Bie. Hit song science once again a science? *Proceedings of the 4th International Workshop on Machine Learning and Music: Learning from Musical Structure*, 2011.
- [64] Eric Nichols, Dan Morris, Sumit Basu, and Christopher Raphael. Relationships between lyrics and melody in popular music. In *International Society on Music Information Retrieval*, pages 471–476, 2009.
- [65] Jyrki Niemi, Krister Lindén, and Mirka Hyvärinen. Using a bilingual resource to add synonyms to a Wordnet: Finnwordnet and Wikipedia as an example. *Global WordNet Association*, pages 227–231, 2012.
- [66] Boulevard of Broken Dreams Lyrics Video. https://www.youtube.com/results?search_query=boulevard+of+broken+dreams+lyrics.
- [67] Shazam Hall of Fame. <http://www.shazam.com/hall-of-fame>.
- [68] François Pachet and Pierre Roy. Hit song science is not yet a science. In *International Society for Music Information Retrieval*, pages 355–360, 2008.
- [69] Giuseppe Pirró and Jérôme Euzenat. A feature and information theoretic framework for semantic similarity and relatedness. In *International Semantic Web Conference*, pages 615–630, 2010.
- [70] The Difference Between Poetry and Song Lyrics. <http://bostonreview.net/forum/poetry-brink/difference-between-poetry-and-song-lyrics>.

- [71] Rhyme Zone. <http://rhymezone.com/>.
- [72] Rowan Righelato. Bill Callahan doesn't just write songs, he sings poems. *The Guardian*, 19 September 2013.
- [73] Robin Frederick. To rhyme or not to rhyme, <http://mysongcoach.com/rhyming-in-contemporary-songs/>. 2013.
- [74] Peter Mark Roget. *Roget's Thesaurus of English Words and Phrases*. TY Crowell Company, 1911.
- [75] Roy Shuker. *Understanding popular music*. Psychology Press, 1994.
- [76] Abhishek Singhi and Daniel G Brown. Are poetry and lyrics all that different? In *International Society for Music Information Retrieval*, pages 471–476, 2014.
- [77] Abhishek Singhi and Daniel G Brown. Hit song detection using lyric features alone. In *International Society for Music Information Retrieval*, 2014.
- [78] Abhishek Singhi and Daniel G Brown. On cultural, textual and experiential aspects of music mood. In *International Society for Music Information Retrieval*, pages 1–6, 2014.
- [79] Abhishek Singhi and Daniel G Brown. Can song lyrics predict hits? *To appear in International Symposium on Computer Music Multidisciplinary Research*, 2015.
- [80] Alex G Smith, Christopher XS Zee, and Alexandra L Uitdenbogerd. In your eyes: Identifying clichés in song lyrics. In *The Australasian Language Technology Association Workshop*, pages 88–96, 2012.
- [81] Stephen Sondheim. *Finishing the hat: collected lyrics (1954-1981) with attendant comments, principles, heresies, grudges, whines and anecdotes*. Knopf, 2010.
- [82] Stephen Sondheim. *Look, I made a hat: collected lyrics (1981-2011), with attendant comments, amplifications, dogmas, harangues, digressions, anecdotes and miscellany*. Knopf, 2011.
- [83] Dana Stevens and Francine Prose. Bob dylan: Musician or poet? *The New York Times*, page BR27, 17 December 2013.
- [84] Valerie N Stratton and Annette H Zalanowski. Affective impact of music vs. lyrics. *Empirical Studies of the Arts*, 12(2):173–184, 1994.

- [85] Steven Suskin. *Opening Night on Broadway: A Critical Quotebook of the Golden Era of the Musical Theatre, Oklahoma!(1943) to Fiddler on the Roof (1964)*. Schirmer Trade Books, 1990.
- [86] WikiSynonyms: Find synonyms using Wikipedia redirects. <http://wikisynonyms.ipeirotis.com/search>.
- [87] Online Thesaurus. <http://www.merriam-webster.com>.
- [88] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *International Society for Music Information Retrieval*, pages 325–330, 2008.
- [89] Dirk Vanderbeke. Rhymes without reason? or: The improbable evolution of poetry. *Politics and Culture*, 1, 2010.
- [90] Villarreal, Alexandra. 5 Reasons Every Hopeless Romantic Needs to Listen to Ed Sheeran. <http://mic.com/articles/73449/5-reasons-every-hopeless-romantic-needs-to-listen-to-ed-sheeran>. 2013.
- [91] Lyric Writing vs. Poetry. <http://www.writersdigest.com/qp7-migration-books/writing-better-lyrics-interview>.
- [92] Wikipedia. <http://en.wikipedia.org/>.
- [93] Hua Wu and Ming Zhou. Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing*, volume 16, pages 72–79, 2003.
- [94] Yi-Hsuan Yang and Homer H Chen. *Music emotion recognition*. CRC Press, 2011.
- [95] Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. Toward multi-modal music emotion classification. In *Pacific Rim Conference on Multimedia*, pages 70–79, 2008.
- [96] William Butler Yeats. *The wild swans at Coole*. Macmillan, 1919.