Can Song Lyrics Predict Hits?

Abhishek Singhi and Daniel G. Brown

University of Waterloo,
Cheriton School of Computer Science
{asinghi,dan.brown}@uwaterloo.com

Abstract. The music information retrieval task of predicting hits is largely unsolved. 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 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 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 14 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.

Keywords: hit song detection, rhymes, meter, music information retrieval, bayesian networks, support vector machines.

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 [1], has traditionally been seen as extremely hard [2]. 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. 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 [1, 2], though Dhanaraj and Logan [3] used both text and audio features in their experiments.

Our focus in this paper 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 [4], 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, and offer songwrites an 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 [5].

We make use of song lyrics to build our models, though in later experiments we added 14 audio features from Echo Nest [6] 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 [7], 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 2.

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 2014. 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 [8] and Bayesian network classifiers from Weka [9], with 10-fold cross validation. The SVM slightly outperforms the Bayesian network, but we focus our presentation on the 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 songwriting.

2 Related Work

Several authors have previously attempted to predict hit songs, largely using audio features. Dhanaraj and Logan [3] 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. A key limitation of their work is that they used features designed for prose, rather than ones designed for verse. Ni et al. [1] 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 [2] 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 [10] 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. [11] 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. [12] 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. [13] 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. [13], 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 which 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 [3], we select hits from the shorter 2008-2013 interval, as music taste evolves over time. Also, our data sets are available at `www.cs.uwaterloo.ca/~browndg/CMMR15data`.

3 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.

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 [1] 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 1.

Table 1: The number of hits and flops in our data set using the four definitions of flops.

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

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. 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.

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 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. 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 [14].

For all songs, we obtained lyrics from a free online lyrics repository [15]. 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 [16], while the list of single release of artists were obtained from the artists' discography pages on Wikipedia [17].

4 Method

We use the complete set of 24 rhyme and syllable features of the Rhyme Analyzer [7], 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 Table 3.

Table 2: The list of lyric features used by our algorithm.

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

We refer the reader to Hirjee and Brown [18] for more information about these features.

Table 3: Definitions of different kinds of internal rhymes.

Internal Rhyme	**Definition**
Chain rhymes	Consecutive words or phrases in which each rhymes with the previous
Compound rhymes	Two pairs of line-internal rhymes overlapping within a single line
Bridge rhymes	Internal rhymes spanning two lines where both the members are internal
Link rhymes	A rhyme between the end of one line and an internal part of the next

We use a total of 31 lyrics features, which are defined in Table 2. 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 [19] 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 4. We use a total of 31 lyric features, the definition of which can be found in Table 2. 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 craftmanship. We use different definitions of flops, as described in the previous section, and include no flops shorter than 30 lines.

Table 4: Stress pattern in different meter styles.

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

Most previous works in hit detection [1, 2, 11] have used audio features to discern hits from flops. Some, like Dhanaraj and Logan [3], 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 [20]. We obtained these audio features via the Echo Nest's APIs [6]. We discarded the songs for which we could not find the audio features from Echo Nest [6]; 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 [8], 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. We also used Bayesian network module from Weka [9] 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.

5 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 and 6.

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 3. 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.

5.1 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 and 6 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 and the receiver operating characteristic curve plot is the left curve in Figure 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 AUC because a high-AUC classifier is not always better than a low-AUC one [21] and the matrices give a more complete picture.

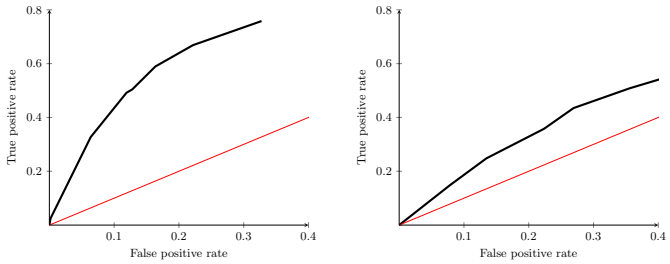

Fig. 1: The receiver operating characteristic curve obtained when using the first and second definition of flops respectively. The area under the ROC curve is 0.688 and 0.573 respectively.

Table 5: 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	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

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 much better than audio features in predicting hits, and that adding audio features only slightly improved the results. The resuts obtained using a Bayesian network is in Table 7 and the ROC curve is depicted in Figure 2. The confusion matrices obtained using both the audio and lyrics features and a weighted-cost SVM are in Table 8.

Table 6: The results we obtain using a Bayesian network.

	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 7: Lyrics features outperforms the audio features in discering hits from flops.

	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

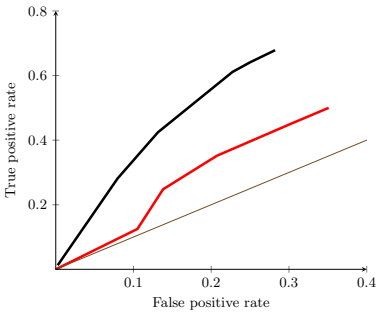

Fig. 2: The ROC curve obtained when using the first definition of flop and a weighted-cost SVM. The black and the red curves are obtained using the lyrics and audio features respectively. The AUC using the lyrics and audio features is 0.692 and 0.572 respectively.

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 verify 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 9. 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.

Table 8: Surprisingly, the naïve Bayesian network gives us better result than weighted-cost SVM when using both audio and lyrics features.

		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 9: We see a noticable improvement in performance with lyrics longer than 50 lines, which are more accurate than the shorter ones.

		Predicted Value
True 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

5.2 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 and 6 respectively, choosing a weight for the SVM that gives a recall $\approx 50\%$, and the ROC plot is the right curve in Figure 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.

5.3 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 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 and 6 respectively. Surprisingly, using Bayesian network we obtain a trivial classifier, which may be because of the small data set. 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.

5.4 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 flies high. 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 and 6 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.

6 What makes a song a hit?

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 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 10 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.

Table 10: The outcome of our algorithm on hits in our data set using the first definition of flops for four popular artist.

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

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

7 Conclusion

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 Bayesian network does not perform well on smaller data sets and is usually outperformed by weighted-cost SVM. The motivation for using Bayesian network is to obtain the probability distribution of the features over the hits and the flops to identify features playing a vital role in determining a hit. We see that the presence of lots of 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 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. An obvious drawback of this approach is that we cannot predict if the outcome of a cover/remake of a song is going to be any different from the original song.

Our work is novel and simple, and it outperforms previous hit detection models. We will provide the data used at `www.cs.uwaterloo.ca/~browndg/CMMR15data` to enable researchers to reproduce and improve upon this work. An extension might be to combine these features with features derived from either recordings, scores or text complexity, and to focus on specific genres.

References

1. Ni, Y., Santos-Rodríguez, R., McVicar, M., De Bie, T.: Hit song science once again a science? Proceedings of the 4th International Workshop on Machine Learning and Music: Learning from Musical Structure (2011)
2. Pachet, F., Roy, P.: Hit Song Science Is Not Yet a Science. In: International Society for Music Information Retrieval. pp. 355–360 (2008)
3. Dhanaraj, R., Logan, B.: Automatic Prediction of Hit Songs. In: International Society for Music Information Retrieval. pp. 488–491 (2005)
4. Anderson, C.A., Carnagey, N.L., Eubanks, J.: 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)

5. Henard, D.H., Rossetti, C.L.: All you need is love? Communication insights from pop music's number-one hits. Journal of Advertising Research 54(2), 178–191 (2014)
6. Echo Nest, http://echonest.com/
7. Hirjee, H., Brown, D.G.: Rhyme Analyzer: An analysis tool for rap lyrics. In: International Society for Music Information Retrieval (2010)
8. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. In: ACM SIGKDD Explorations Newsletter. vol. 11, pp. 10–18 (2009)
10. Fan, J., Casey, M.A.: Study of Chinese and UK hit Songs Prediction. Proceedings of Computer Music Multidisciplinary Research (CMMR) (2013)
11. Herremans, D., Martens, D., Sörensen, K.: Dance hit song prediction. In: International Workshop on Machine Learning and Music, ECML/PKDD (2013)
12. Bischoff, K., Firan, C.S., Georgescu, M., Nejdl, W., Paiu, R.: Social knowledge-driven music hit prediction. In: Proceedings of the Advanced Data Mining and Applications, pp. 43–54 (2009)
13. Smith, A.G., Zee, C.X., Uitdenbogerd, A.L.: In your eyes: Identifying clichés in song lyrics. In: Proceedings of the Australasian Language Technology Association Workshop. pp. 88–96 (2012)
14. Shuker, R.: Understanding popular music. Psychology Press (1994)
15. Metro Lyrics, http://www.metrolyrics.com/
16. Billboard, http://www.billboard.com/
17. Wikipedia, http://en.wikipedia.org/
18. Hirjee, H., Brown, D.G.: Using automated rhyme detection to characterize rhyming style in rap music. Empirical Musicology Review 5(4), 121–145 (2010)
19. Elovitz, H., Johnson, R., McHugh, A., Shore, J.: Letter-to-sound rules for automatic translation of English text to phonetics. In: IEEE Transactions on Acoustics, Speech and Signal Processing. vol. 24, pp. 446–459 (1976)
20. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: International Society for Music Information Retrieval. pp. 591–596 (2011)
21. Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters 27(8), 861–874 (2006)