Using Listeners' Interpretations in Topic Classification of Song Lyrics

Varvara Papazoglou **Robert Gaizauskas**
Department of Computer Science, University of Sheffield
{vpapazoglou1, r.gaizauskas}@sheffield.ac.uk

Abstract

Incorporating listeners' interpretations of song lyrics has been shown to significantly improve topic classification accuracy. Using a different type of interpretation, as compared to previous research, we propose four possible representations of songs as input for classification systems. The results show that (a) some representations are consistently better than others, and (b) the similarity of topic classes along with the ambiguity of song lyrics may affect the classification accuracy, which argues for using top-n classification (n>1) and associating multiple top ranking classes with each song. We also examine the case of training a system on both lyrics and interpretations and testing it on songs that lack interpretations.

1 Introduction

Song lyrics differ from prose text in various ways: they tend to be more ambiguous, to contain more figures of speech, to break syntactic rules, to be accompanied by music and to have a rhythm. Considering that the majority of popular music contains lyrics, it is assumed that a lot of information about songs can be extracted from lyrics. This information can be useful for many Music Information Retrieval (MIR) tasks, such as music recommendation, classification, and search.

We focus on the task of automatic topic classification of English-language songs based on song lyrics and interpretations of them. The interpretations have been retrieved from a website that hosts song lyrics and interpretations generated by the website's users. Our approach is novel in that (a) opposed to previous research, the interpretations we use refer to specific fragments of lyrics and not to the whole song (so it is less probable that they contain information unrelated to the topic of the song), (b) we propose a novel representation of songs using lyrics and interpretations, which consistently achieves high classification accuracy, (c) we examine a top-n topic classification approach, and (d) we combine lyrics and interpretations in an attempt to improve classification of unseen lyrics for which there are no available interpretations (e.g., recently released songs).

Our aim is to investigate what is the best song representation for the task and to create a system which predicts topics that meet listeners' needs and expectations. Our main hypotheses are that interpretations are more informative than lyrics in determining the topic of a song, and that a top-n topic classification approach is useful from the users' perspective. The intuition for the latter is that a song can actually belong to more than one topic class either because some topic classes are semantically related or because the song has indeed more than one topic. To illustrate our point, let us consider the case of a song A which talks about a breakup and heartache, and a song B which talks about a breakup but not heartache Although we could initially consider these two topics to be similar, there are cases of songs which do not belong to both topics. We assume that a user who searches for music based on the topic of the lyrics would be satisfied if multiple related topics were assigned to a song instead of a single one.

Our results suggest that listeners' interpretations of lyrics indeed improve the accuracy of the classification, and that top-n classification is an effective approach. However, using lyrics and their interpretations in the training stage and lyrics in the test stage does not significantly and consistently improve accuracy compared to using solely lyrics for both stages.

2 Related Work

Lyrics have been used in a range of MIR tasks, sometimes combined with acoustic properties of the respective songs. Watanabe and Goto (2020) introduce Lyrics Information Processing (LIP) as a research field specific to analysis and generation of lyrics, and present a range of applications. It is

Song Topics			
Sex	317	Political statement	191
Heartache	294	Death	190
Girl	277	War	185
Religion	272	Events in the news	179
Drugs	265	Cheating	158
Ex-partner	240	Dealing with fame	156
Parent	208	Autobiographical	154
Dead friend	205	Depression	154
Places	203	Criminals	147
Breakup	199	Loneliness/isolation	147

Table 1: Number of songs per topic in the dataset (prior to splitting into training and test set).

also worth mentioning that users of music search systems appear to use lyrics frequently (Lee and Downie, 2004).

Some of the first approaches to topic detection of song lyrics use clustering methods. Kleedorfer et al. (2008) used Non-Negative Matrix Factorisation; Sasaki et al. (2014) used Latent Dirichlet Allocation in order to detect and visualise five of the latent topics of the lyrics in an interactive system.

More recent research exploits listeners' interpretations of lyrics for topic classification (Choi, 2018; Choi and Downie, 2018; Choi et al., 2016, 2014). The highest accuracy was achieved when interpretations or the concatenation of lyrics and interpretations were used as features instead of the lyrics alone.

3 Data

Song topics and song titles are collected from Songfacts[1]. Songfacts provides information about songs and artists and assigns categories to the songs manually, based on sources like interviews, publicity releases, press, etc. We collect all the song titles and topics from the category "about", which contains 206 topics. There is no hierarchy in the topics, and some songs belong to more than one topic.

These song titles are then searched for in Genius[2], from where their lyrics and their interpretations are collected. In Genius, users annotate specific fragments of lyrics (e.g., one or more consecutive words or lines) with an interpretation. The users can upvote and downvote the suggested annotations, so the final interpretations usually reflect the single most widely acceptable view on the meaning of the song.

[1]https://www.songfacts.com
[2]https://genius.com

We selected the 20 most populated topics for our dataset (Table 1). The intuition is that in order to meet listeners' needs, the system should cover a relatively large number of topics, while at the same time guarantee that there are enough songs per topic for training the classifier. In this set the vast majority of songs belong to a single topic. The few songs left belonging to multiple topics are then assigned to the less populated of the 20 topics. Prior to this, we ensure that the language of each selected song's lyrics is English, using the Python module langdetect[3], a port of a library by Nakatani (2010). The final training dataset is balanced, consisting of 20 topics (130 songs each) and a total number of 2,600 songs, and we also have an unbalanced test set with 1,541 songs. We represent each song in four ways:

1. **Lyrics**: only the lyrics of the song (without the song title).

2. **Interpretations**: concatenation of all interpretations of the lyrics. If the annotated lyric fragments are repeated, the respective interpretations are repeated as well.

3. **Mixed**: starting with the lyrics, we detect fragments which have been annotated with an interpretation and replace these fragments with their respective interpretations. The rest of the lyrics remain unchanged (repetitions are preserved).

4. **Concatenation** of the first two representations.

All text is lowercased, contractions are expanded using the Python module contractions[4], song structure annotations (e.g.: "[Chorus]") are removed, and lemmatisation (WordNet lemmatiser) and stemming (Porter stemmer) are performed, using the Python module NLTK[5].

4 Experimental Setup

Using the scikit-learn Python library[6], we use TFIDF scores of unigrams as features for each of our four song representations. Unigrams with document frequency less than 5 are discarded. Using 5-fold stratified cross-validation we train each

[3]https://pypi.org/project/langdetect
[4]https://pypi.org/project/contractions
[5]https://www.nltk.org
[6]https://scikit-learn.org version 0.24.1

	Lyrics		Interpretations		Mixed		Concatenation		Concat. & Lyrics	
	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-3
kNN	0.1901	0.3790	0.2842	0.4822	0.2862	0.4646	0.2706	0.4763	0.1707	0.3310
LR	**0.3348**	**0.5892**	**0.4549**	**0.6788**	**0.4692**	**0.7333**	**0.4802**	**0.7352**	**0.3355**	**0.5964**
MNB	0.2732	0.5204	0.3180	0.5626	0.3731	0.6230	0.3679	0.6178	0.2726	0.5088
RF	0.2330	0.4276	0.3783	0.5912	0.4114	0.6424	0.3855	0.6275	0.2524	0.4640

Table 2: Accuracy scores for top-1 and top-3 classification with N=20 topic classes.

of four classification algorithms (described in the next paragraph) on each representation. In the top-1 classification approach, we consider the predicted topic to be the one that the classifier predicts. In the top-n classification approach for n>1, during the testing stage each of the classifiers returns the predicted probabilities per topic class for the current song. If the true topic class is one of the top-n predicted topic classes, then we consider the song to have been classified correctly; otherwise we consider the class with the highest probability to be the predicted class and the song to have been classified incorrectly. The intuition is that a song with a true label A but predicted label B should not be considered misclassified if A and B are in the top-n predicted classes. Besides, songs can be interpreted in different ways, so returning a small number of possible topics to a user who searches for songs on a specific topic is acceptable. Moreover, this approach covers cases of topics that are semantically similar to each other (if A and B are semantically similar, then predicting B should not be considered incorrect). Since we have a dataset with N=20 classes, we have selected n=3. For smaller N values we prefer decreasing n as well (e.g., for N=10, n=2 is intuitively more appropriate).

We have experimented with the following classification algorithms[7] (their parameters were selected using grid search): k-Nearest Neighbours (kNN, n_neighbors=5, weights='distance'), Logistic Regression (LR, random_state=17, max_iter=1000), Multinomial Naïve Bayes (MNB, default parameters), Random Forest (RF, random_state=17, class_weight='balanced', criterion='gini')

We also perform two classification experiments: in the first experiment, the training and test stages use features of the same song representation (i.e., either lyrics or interpretations or mixed representation or concatenation), while in the second experiment the training stage uses the mixed concatenation of lyrics and interpretations, and the test stage uses only lyrics. This is to test the hypothesis that training on lyrics and interpretations will lead to better classification of new songs without accompanying interpretations. We use the same training and test sets for both experiments.

[7]From scikit-learn.

5 Results

Table 2 contains the accuracy scores for each algorithm for both experiments. LR consistently returns the highest accuracy for all settings. For the first experiment, the two representations that return the highest accuracy are the mixed lyrics-interpretations representation and the concatenation of lyrics and interpretations. When only lyrics are used, the accuracy is consistently lower.

Training each classifier on the concatenation of lyrics and interpretations and testing on lyrics compared to training solely on lyrics (second experiment, last column in Table 2) improves results significantly with RF and marginally with LR, while it actually reduces accuracy scores with kNN and MNB.

6 Discussion

Results do not support the hypothesis that training on lyrics and interpretations will improve classification of unseen lyrics without interpretations. However, this does not necessarily imply that combining lyrics with interpretations is not helpful in improving classification of song lyrics. It is possible that better feature engineering and preprocessing might actually make this approach very effective.

Comparing the results between top-1 and top-3 classification approaches we noticed that there are indeed some frequently confused topic classes, such as: (a) events in the news, political statements, war, (b) heartache, breakup, ex-partner, cheating. In both examples, the classes seem to be

	Lyrics		Interpretations		Mixed		Concatenation		Concat. & Lyrics	
	top-1	top-2	top-1	top-2	top-1	top-2	top-1	top-2	top-1	top-2
kNN	0.3082	0.6123	0.4073	0.6974	0.4184	0.7322	0.4254	0.7280	0.3040	0.5858
LR	**0.4784**	**0.8020**	**0.5593**	**0.8466**	**0.6318**	**0.9038**	**0.6346**	**0.9052**	**0.4909**	**0.7894**
MNB	0.4198	0.7378	0.4658	0.7922	0.5467	0.8131	0.5467	0.8020	0.4561	0.6960
RF	0.4114	0.7308	0.5216	0.8089	0.5858	0.8452	0.5635	0.8466	0.4059	0.7075

Table 3: Accuracy scores for top-1 and top-2 classification with N=8 topic classes.

similar to each other. This suggests that topic similarity should be taken into account in our dataset.

A comparison of our results to previous research is very useful in order to evaluate our approach. In previous research (Choi, 2018; Choi and Downie, 2018; Choi et al., 2016, 2014), interpretations are in the form of general comments about the lyrics of the whole song and frequently contain information other than the meaning of the lyrics (e.g., how much the particular listener likes the song and why, what it reminds them of, comments about the album or a live concert in which the song was played or a music video, etc.). Both Choi's and our research retrieve song titles and topics from Songfacts. A direct comparison between the results of Choi's and our research is difficult, as we cannot use the same songs mostly due to the availability of interpretations and the fact that we cannot obtain the same dataset from Songfacts, which is updated regularly with new songs and information. However, we try to follow similar preprocessing and feature extraction steps, with the difference that we do not eliminate stopwords; the use of TFIDF weighting lowers the impact of terms with very high frequency in the dataset, so that using a list of standard and corpus-specific stopwords is not required, and in our experiments we did not notice any significant difference with stopword removal. Choi et al. (2016) use a dataset of 800 songs and 8 balanced topic classes that consists of lyrics, interpretations, and concatenation of them. Then, using TFIDF features, they compare four classifiers: kNN, SVM with a linear kernel, SVM with radial basis function kernel, and Naïve Bayes. The highest accuracy score (0.66) is achieved by Linear SVM, using the concatenation of lyrics and interpretations. Interpretations and concatenation consistently return higher accuracy than lyrics. Using fasttext[8] word embeddings and Naïve Bayes on the same dataset, concatenation returns again the highest accuracy (0.5788) (Choi, 2018). Table 3 shows the accuracy scores we achieve with the same four classifiers using our four representations, on the top 8 balanced topic classes (training set: 1,440 songs, test set: 717 songs). For top-1 classification, the highest accuracy score is 0.6346 using concatenation with LR, which appears to be similar with the results achieved in Choi et al. (2016). Using top-2 classification, the accuracy score is significantly improved, reaching 0.9052. For training on concatenation and testing on lyrics, accuracy scores follow a different trend than with N=20, but are still low. Finally, we preferred to use top-2 instead of top-3, due to the small number of topic classes.

[8]https://fasttext.cc

7 Conclusion and Future Work

Our results suggest that the interpretations of the lyrics are indeed more informative than lyrics alone for identifying the topic of the lyrics, which is in line with previous research. The main differences compared to previous research are that: (a) we use interpretations targeted on specific fragments of lyrics instead of interpretations which are in the form of general comments about the lyrics, (b) we examine the impact that training a model with lyrics and their interpretations has on predicting the topic class of unseen song that lack interpretations, and (c) we allow for more flexibility in classification by accepting as correctly classified the songs which have the correct topic class as one of their top-n predicted topic classes. The latter is a reasonable approach for MIR applications which return to the user a list of songs of a selected topic or predict the topic of a specific song.

Using interpretations in the form of comments on specific fragments of lyrics allows us to analyse song lyrics in detail. We are planning to study the possible different impact of chorus and verse terms in topic classification, as well as to experiment with features other than TFIDF unigrams, e.g. word embeddings, and to examine human performance as an evaluation of our approach.

References

Kahyun Choi. 2018. *Computational lyricology: quantitative approaches to understanding song lyrics and their interpretations*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Kahyun Choi and J. Stephen Downie. 2018. Exploratory investigation of word embedding in song lyric topic classification: Promising preliminary results. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 327–328, New York, NY, USA. Association for Computing Machinery.

Kahyun Choi, Jin Ha Lee, and J. Stephen Downie. 2014. What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 453–454.

Kahyun Choi, Jin Ha Lee, Xiao Hu, and J. Stephen Downie. 2016. Music subject classification based on lyrics and user interpretations. *Proceedings of the Association for Information Science and Technology*, 53(1):1–10.

Florian Kleedorfer, Peter Knees, and Tim Pohle. 2008. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval, ISMIR 2008*, pages 287–292.

Jin Ha Lee and J. Stephen Downie. 2004. Survey of music information needs, uses, and seeking behaviours: preliminary findings. In *Proceedings of the 5th International Conference on Music Information Retrieval, ISMIR 2004*.

Shuyo Nakatani. 2010. Language detection library for java.

Shoto Sasaki, Kazuyoshi Yoshii, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima. 2014. Lyricsradar: A lyrics retrieval system based on latent topics of lyrics. In *Proceedings of the 15th International Conference on Music Information Retrieval, ISMIR 2014*, pages 585–590.

Kento Watanabe and Masataka Goto. 2020. Lyrics information processing: Analysis, generation, and applications. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 6–12, Online. Association for Computational Linguistics.