Lyrics Mining for Music Meta-Data Estimation

Hasan Oğul, Başar Kirmaci

Lyrics mining for music meta-data estimation

Hasan Oğul[1,*], Başar Kırmacı[1]

[1]Department of Computer Engineering, Baskent University, Ankara, Turkey
* Corresponding author : hogul@baskent.edu.tr

Abstract. Music meta-data comprise a number of structured attributes that provide descriptive annotations such as singer, author, genre and date of a song deposited in a digital library. While they provide a crucial knowledge to represent music entry in current information retrieval and recommendation systems applications, they suffer from two limitations in practice. First, they may contain missing or wrong attributes due to incomplete submissions. Second, available attributes may not suffice to characterize the music entry for the objective of the retrieval or recommendation task being considered. Here, we offer an automated way of estimating the meta-data of a song using its lyrics content. We focus on attributing the author, genre and release date of songs solely based on the lyrics information. To this end, we introduce a complete text classification framework which takes raw lyrics data as input and report estimated meta-data attributes. The performance of the system is evaluated based on its retrieval ability on a large dataset of Turkish songs, which was gathered in this study and made publicly available. The results promote the use of such technique as a complementary tool in organizing music repositories and implementing music information retrieval systems.

Keywords: Music information retrieval; song classification; author attribution.

1 Introduction

We have witnessed a drastic shift in promotion trends of music industry in response to changes in the way of people's music listening habits in recent years. Collective online shops and libraries have been more popular compared to individual album records. Consequently, the amount of music data available online has increased dramatically over the past few years. There is a pressing need to develop intelligent tools which will improve the usability of these data such that users can access, enjoy and communicate the content in a more effective and flexible way. This need has been further emphasized by the contribution of mobile devices in accessing online music content. The research effort into music information retrieval and recommendation systems has been therefore raised significantly to respond these challenges in the last decade (Orio, 2006; Schedl et al., 2014).

Music is a multimodal object: it comprises an audio signal, song lyrics and other textual annotations that provide descriptive information about the entry, including e.g. song's singer, composer, author, genre, release date and social data. The latter is usu-

adfa, p. 1, 2011.

ally referred as meta-data and constitutes a brief yet helpful representation of music entry for accessing, searching or organizing the content in online media. Music information retrieval is a fruitful effort in more enjoyable access of music content in digital libraries. It usually requires an abstraction technique to represent the music objects and a model of comparing them to retrieve relevant music entries in available repositories. This abstraction task for music objects has been usually approached with meta-data or audio content (Moutselakis and Karakos, 2009; Debaecker eta al., 2011). There are two major practical limitations with using meta-data approach. First, some attributes in the collection might be missing or incorrectly entered by database administrator or submitting user. Second, available attributes may not suffice to characterize the music entry for the objective of the task being considered. For example, if existing meta-data structure does not provide any attribute about the release date of a song, it is useless when a user wishes to retrieve similar songs from a particular period of time. Employing audio content is another widespread approach in extracting features for abstract representation of music (Costa et al., 2004; Sarkar and Saha, 2015). Even though audio is the core component of a music entry, it has several drawbacks to serve as a feature generator to encode the music object. It contains a variety of signal layers corresponding to for example each instrumental sound, voice of a singer and background noise. Therefore, mining an audio content is a difficult task. There is indeed no well-established way of feature representation for a song from its audio content (Schedl et al., 2014).

Although general music perception is largely characterized by melodic and acoustic components comprised by audio content, the overall perception of non-instrumental song in fact can be elucidated only by considering all modalities including its lyrics. In spite of its great potential to characterize several cognitive concepts, research efforts on lyric-based music information retrieval and classification are very few. It was hypothesized that lyrics may contain lexical components that emphasize a certain mood and as such can be used to recognize the underlying mood (van Zaannen and Kanters, 2010). In fact, the words such as "happy", "angry", "smile" and "dead" do not have to be spelled with a strong emotional voice or melody. In this respect, lyrics were used to classify the songs based on several mood categories such as "happy", "sad", "depressed" and "desire" (Hu et al, 2009). In some studies, a similar attempt was referred as lyric-based song sentiment classification, which seeks to assign songs appropriate sentiment labels such as light-hearted and heavy-hearted (Xia et al., 2008). Genre of a song was shown to be predictable by lyrics content (Mayer et al., 2008; Fell and Sporleder, 2014). A feasibility study was presented to recognize the genre of a song from its lyrics written in Nordic language (Adriano et al., 2014), which is, to our knowledge, the only study that considers lyrics-based song classification in a language other than English.

In this study, we focus on estimating a set of meta-data attributes, i.e. author, genre and release date, solely from song lyrics to address the aforementioned challenges. There are a few attempts for genre or mood classification from lyrics in the literature. However, this study, to our knowledge, is the first attempt to predict the author and release date of a song from its lyrics, where some preliminary results have been presented in Kirmaci and Ogul (2015). We offer a number of novel features that are believed to be representative for song lyrics. We present the results of a comprehensive analysis conducted on a large dataset of Turkish songs, which was also collected in this study, on a rigorous experimental setup. Experimental results suggest the use of proposed technique as a complementary tool in music information retrieval applications.

2 Approach

2.1 Classification framework

The task is to assign unknown lyrics content into known classes of different attributes. Here, we consider three meta-data attributes for a song: author, genre and release date. For each attribute, classes are restricted to a set of known labels. In a supervised approach, the framework learns a model to distinguish between classes for an attribute. This is achieved by feeding a learning classifier by a set of numeric values that represent the content of song objects in the training collection. Having a trained model, the prediction stage involves feeding the learned model with the representation of query song, which is encoded in the same way. The result is an estimation of class among the restricted label set of that attribute. There are two issues that should be dealt with in this framework: (1) which classification model will be learned, and (2) which features will be used to feed that classifier. Pertaining to its prior success in similar applications, we opted to use here a Naive Bayes classifier with a multinomial assumption in the distribution of data (Kibriya, 2004).

Naive Bayes is a supervised classification technique based on Bayesian statistics. Bayesian statistics approach assumes an underlying probabilistic model and it allows user to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. Parameter estimation for naive Bayes models is done using maximum likelihood. In Multinomial Naïve Bayes (MNB), a multinomial probability distribution of data is assumed.

Given a set of variables, $X = \{x_1,x_2,...,x_d\}$, it is desired to construct the posterior probability for the class C_j among a set of possible outcomes $C = \{c_1,c_2,...c_d\}$. In our case, X is the vector of features derived from the lyrics content and C is the set of classes present in the meta-data attribute being studied. Using Bayes' rule:

$$p(C_j \mid x_1, x_2,..., x_d) = p(x_1, x_2,..., x_d \mid C_j) p(C_j)$$

where $p(Cj \mid x_1,x_2,...,x_d)$ is the posterior probability of class membership, i.e., the probability that the song X belongs to C_j. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood to a product of terms:

$$p(X \mid C_j) = \prod_{k=1}^{d} p(x_k \mid C_j)$$

and rewrite the posterior as:

$$p(C_j \mid X) = p(C_j)\prod_{k=1}^{d} p(x_k \mid C_j)$$

Using Bayes' rule above, we label a new song *X* with a class label C_j that achieves the highest posterior probability. Although the assumption that the variables are independent is not always accurate, it can significantly simplify the classification task, since it allows the class conditional densities $p(x_k|C_j)$ to be calculated separately for each variable, i.e., it reduces a multidimensional task to a number of one-dimensional ones. Furthermore, the assumption does not largely affect the posterior probabilities, thus, it leaves the classification task unaffected.

2.2 Lyrics Features

A major concern in text classification is how to select the numerical features to be derived from lyrics text to represent the category that we seek to assign the song being queried. In fact, same representation is needed for all samples in a collection which will be used to train a supervised model. Since a song lyrics has a different characteristic than text content of a scientific article, a blog entry or a magazine news, typical feature representations used in previous text classification applications formerly discussed needs to be re-evaluated for the objective of song classification by text. Since lyrics may exhibit certain structures due to its specific parts such as chorus, bridge and verse, other features might be more relevant in retrieval studies.

In this study, we consider five feature sets applicable to song lyrics: bag-of-words, word N-grams, character N-grams, global text statistics and line length statistics. Simply concatenating some of these feature sets can provide fusion of several characteristics which might be useful in associating lyrics content with meta-data attributes.

Bag-of-words. Bag-of-words method is a classical yet powerful method for numerical representation of text objects. In this method, each unique word appeared in any document in the collection is considered as a separate feature. To encode a text entry, the information about the presence of each word is quantified to and used to fill out the feature vector indexed by these words. A simple way to quantify a word is to use a Boolean model, which only identifies whether the given word appears at least only once in the current document. Even if it is sometimes useful, a more sophisticated method has been used based on weighting each word and associating it with its frequency in the document. The method is called as term frequency-inverse document frequency (tf-idf), which considers both the frequency of terms in current documents and its occurrence statistics in other documents in the collection. Although using tf-idf has been shown to be more explanatory in pairwise comparison of documents in many cases, we opt to use only term frequency in our framework. The reason behind this choice is the assumption that the machine learning classifier that we use can already handle the statistics of each feature in other samples of the collection. That is, for example, a term will have no effect in separating hyper-plane of the trained model if it is present with a near frequency in all documents of the collection. Another crucial choice in implementing a bag-of-word method is whether to use a stemming for words or not. Stemming implies the preference of the use of original words without considering suffixes and reduces the size of feature space. We hypothesize here the using the frequencies of word roots as separate features but supporting this representation with the use of suffixes will present a more descriptive representation of author preferences. Hence, we opt to apply stemming in our bag-of-words representation. We implement word stemming by relevant tool in Zemberek, which is an open-source NLP library developed for Turkish language (Akin and Akin, 2007).

Word N-grams. A word N-gram is the frequency of a phrase having N consecutive words. Concatenating the frequency of N-grams for all possible such phrases creates a word N-gram model for text representation. It is commonly used in general text classification applications and author attribution studies. Although it usually provides a complementary representation, it exponentially increases the size of feature space in terms of the value of N. Moreover, use of a large value of N makes the feature vector to have a too sparse content and pervert the learning model. Enlarging the dimension of feature vector with a sparse content will indeed increase the need for the number of samples in training data. Therefore, we determine to use only 2-grams to represent the preference of word phrase uses.

Character N-grams. Use of character frequencies in the form of N-grams (Cavnar and Trenkle, 1994) has been shown to be helpful in several contexts (Iliev et al., 2014). A character N-gram refers to the presence of a string with a length of N characters in a text object. Similar to word N-gram model, a feature set of character N - grams contains the frequency of all possible strings having N characters from a finite alphabet. Character N-gram model does not entail that all such strings correspond to meaningful words in the language. Rather than providing a semantic value, it promises to infer lexical, grammatical or orthographic preferences without any linguistic background. In our case, these features may indirectly help in capturing the preferences in the use of similar rhymes in distinct genres or authors, since their frequency will be high. Another implicit value that character N-grams model maintains is the ability to measure the suffix composition in the text. This is particularly useful for agglutinating languages such as Turkish. In our framework, we built a feature set comprising character 2-grams, 3-grams and 4-grams since longer strings will be resulted with sparse and high dimensional vector content which is not usually desirable as discussed for word N-grams.

Global text statistics. In addition to vocabulary and language preferences, some other global indicators might be descriptive in representing lyrics content. In our framework, we consider four features in this set: total word count, total character count, average word length and number of unique words in the text. The first two features measure the tendency of telling something with long or short phrases. Third feature is related to the preference of using sophisticated and unordinary words, which are usually longer than daily conversation terms. Last feature is about the enrichment of vocabulary use.

Line length statistics. While previous feature sets are application to other text classification studies, we offer some new features only relevant to song lyrics. Lyrics have a certain characteristic of poetical design over a rhythm and melody. This leads to particular layout of lines and stanzas over the entire text, which is not observed in plain texts. Our observation is that the characteristics of this design exhibit a fronting diversity over authors, genres and even the periods of time when the song is released.

Therefore, we include three novel features in this set: average line length, standard deviation of line length and difference between lengths of longest and shortest lines.

3 Results

3.1 Data

We gathered 1048 Turkish song lyrics of 12 authors who have distinct styles. Each author often writes lyrics for the songs composed in one of three different genres: pop, rock and arabesk. The genres were selected based on their popularity in Turkey (Angi, 2013). The authors included in the dataset are listed in Table 1 with the number of their song lyrics in the collection. In the resulting set, the number of authors in each genre is equal, and the number of samples is very close to each other in all categories. When assigning the songs into classes based on their release dates, we concerned with the issue of sociocultural changes in the country which are relevant to corresponding dates. This consideration resulted with the following three classes for the songs: released before 1993, released between 1994 and 2006, released after 2007. The number of samples in each class then became 293, 562, 193 respectively. The year 1993 is the date that Turkish pop music had a great explosion and highly popularized in comparison to arabesk music in especially urban populations. This resulted with a free-style content in both melody and lyrics compared to more conservative and artistic trends in older productions. In mid of 2000s, several rock groups appeared in the country, which significantly affect the listening styles of new generation. The popularity of rock music affected the lyric styles in general. Instead of using words and phrases from daily conversations, songwriters were being enforced to creative linguistic designs in their lyrics.

Table 1. Data set content.

Author	Genre	Number of song lyrics
Sezen Aksu	Pop	124
Serdar Ortaç	Pop	134
Yaşar	Pop	49
Mustafa Sandal	Pop	58
Teoman	Rock	68
Haluk Levent	Rock	57
Barış Manço	Rock	76
Şebnem Ferah	Rock	65
Selami Şahin	Arabesk	92
Yıldız Tilbe	Arabesk	57
Ferdi Tayfur	Arabesk	177
Hakan Altun	Arabesk	91

3.2 Experiment Setup

We conducted classification experiments independently per meta-data category in a ten-fold cross-validation setup. In this setup, the dataset is divided into ten equal partitions such that each partition has a balanced number of samples from all categories. Each sample is then predicted using the classification models trained by other nine partitions which do not have the query sample. All samples are guaranteed to go through a prediction stage after repetition of same experiments ten times with a different training set in each. Proposed classification and feature selection methods were compiled with Weka, an open source data mining tool (Hall et al., 2009). In addition to traditional metrics such as recall, precision and F-measure, we determine the classification performance using ROC (Receiver Operating Characteristic) score. The ROC score is a metric often used to evaluate the performance of information retrieval systems. A score is computed for each category separately by the area under the corresponding ROC curve. Average of these scores is reported as a general ROC score. We used the ROC score to determine the ability of the feature representation schemes to infer the relevance of two songs in terms of meta-data attributes defined.

3.3 Empirical results

We first report the effect of each feature set in meta-data estimation performance. To this end, we compiled MNB to predict author, genre, and release date with a variety of feature set combinations. Pertaining to the ROC scores obtained from cross-validation experiments, the combination of bag-of-words and character N-grams achieved the greatest performance in author prediction (Table 2). In genre and release date prediction, addition of global statistics and line length features improved the overall performance compared with the other combinations of feature sets. According to the results, word-N-grams are not descriptive enough for the song meta-data attributes studied as they can not provide an improvement in ROC scores when combined with bag-of-words features. For that reason, they were not incorporated into further feature set combinations.

Table 2. ROC scores of MNB classifier with varying feature sets with no stemming and no feature selection.

Feature sets					**ROC**		
Bag-of-words	**Word N-Grams**	**Character N-grams**	**Global text statistics**	**Line length statistics**	**Author**	**Genre**	**Date**
√					0,839	0,822	0,650
√	√				0,811	0,820	0,621
√		√			0,862	0,865	0,665
√		√	√		0,856	0,868	0,664
√		√	√	√	0,860	0,873	0,674

To discern the effect of word stemming, we repeated the experiments with the last configuration in Table 2 but now using the words on which a stemming was not applied. Table 3 shows that stemming can promote the estimation performance to some degree. According to the results, author and date predictions can be improved by approximately 2% when stemming is applied, however, genre prediction is not affected by this treatment. This slight improvement in performance indicates that language-specific information, when available, can improve the descriptive ability of representation scheme in retrieval task. In fact, since Turkish is an agglutinating language, stemming can be considered as an informative attempt to characterize the writing style of a Turkish author. On the other hand, positive contribution of character N-grams in prediction performance argues that use of suffixes is also an importance indicator in writing style. This justifies our hypothesis that using the frequencies of word roots as separate features but supporting this representation with the use of suffixes will present a more descriptive representation of author preferences.

Table 3. Effect of stemming in ROC score with best feature set and no feature selection

Category	Without stemming	With stemming
Author	0,841	0,860
Genre	0,875	0,873
Release Date	0,654	0,674

Since the machine learning classifier used for training and prediction can sometimes significantly affect the general performance, we compared the results of multinomial Naive Bayes with some other methods. To justify the assumption of multinomial distribution of data, we repeated the experiments with Gaussian Naive Bayes. We also compiled two different versions of Support Vector Machine (SVM) classifier. SVM is a popular learning method, which has shown to be superior to other algorithms in several classification contexts. A key component in the success of SVM is the kernel function that maps the current data into a higher dimension to search for a hyperplane that can linearly separate the training samples. We opted to use 'linear' and 'radial basis function (RBF)' kernels with their default parameters in our comparative experiments. Table 4 illustrates that multinomial assumption is a reasonable decision since Gaussian NB performed very badly in predicting all attributes. Furthermore, multinomial NB significantly outperformed both SVM-based classifiers in all categorization schemes.

Table 4. Comparing classifiers in terms of ROC score

Category	Naive Bayes (Gaussian)	SVM (RBF)	SVM (Linear)	Naive Bayes (Multinomial)
Author	0,679	0,579	0,736	0,860
Genre	0,687	0,678	0,770	0,873
Release Date	0,578	0,514	0,610	0,674

Overall performance of proposed framework is discerned in Table 5. The table demonstrates the recall, precision, F-measure and average ROC score evaluations for the best classifier configuration for each meta-data attribute. These results suggest that the model for genre classification from lyrics can be used in practical applications as such, pertaining to high scores in both recall and precision. Recall and precision scores for release date prediction are around 70%. As the release date is actually a continuous variable, but here it was enforced to have categorical labels using artificial boundaries, the prediction results for these labels can be considered to be promising. A relatively high ROC score justifies this argument. According to the results, direct prediction of author from lyrics is not reliable enough when we base our assessment on recall and precision scores. On the other hand, a reasonable high ROC achieved in the experiments indicates that the model can successfully rank the candidate songs in relevant to their authors. This suggests that the knowledge contained in feature representation model can provide a good complement to music audio and other modalities in an information retrieval setup when author is an issue of relevance.

Table 5. Estimation performance of NMB with best classification setup

Category	Recall	Precision	F-measure	ROC
Author	0,49	0,60	0,44	0,862
Genre	0,82	0,82	0,82	0,925
Release Date	0,72	0,71	0,71	0,818

We further analyzed the results of best configuration for individual class labels in corresponding meta-data attributes. The results in Table 6 indicate that the authors writing in the same genre category have similar preferences. For example, most confusing author for SA was SO, where both write songs in pop category. This observation was same for all pop music authors. While FT had the highest recall, other authors in Arabesk category were mostly confused with this author.

Table 6. Confusion matrix for author prediction with best classification setup

Predicted	SA	SO	Y	MS	T	HL	BM	ŞF	SŞ	YT	FT	HA
Actual												
Sezen Aksu	78	20	0	0	2	0	5	4	2	0	12	1
Serdar Ortaç	1	110	0	0	1	0	0	1	0	0	19	2
Yaşar	17	10	1	0	1	0	1	0	0	0	18	1
M. Sandal	10	25	0	1	1	0	1	3	0	0	15	2
Teoman	9	7	0	0	47	0	0	3	0	0	2	0
Haluk Levent	17	6	0	0	2	8	4	2	2	0	15	1
Barış Manço	13	9	1	0	1	1	37	2	1	0	11	0
Ş. Ferah	16	10	0	0	5	0	0	30	1	0	3	0
Selami Şahin	6	10	0	0	0	0	1	1	20	0	50	4
Yıldız Tilbe	9	28	0	0	0	1	0	0	1	3	15	0
Ferdi Tayfur	4	9	1	0	0	0	2	0	2	1	158	0
Hakan Altun	3	21	0	0	2	0	0	0	0	0	43	22

According to the confusion matrix for genre classification (Table 7), arabesk is the most distinguishable category. The recall for arabesk classification is computed as 87.5% while it is 76.7% and 78.6 for pop and rock respectively. It is obvious that the lyrics in arabesk songs have special characteristics due to their dispirited emotion, while pop and rock songs may have versatile spiritual characteristics. On the other hand, highest precision was achieved for rock classification with 88.2%. It can also be inferred from the confusion matrix that pop and arabesk songs are usually confused with each other but not with rock songs. This result might be attributable to the fact that rock authors tend toward using an original and more sophisticated vocabulary to create a narrative or legendary effect in their lyrics rather than putting toward a poetic sound. While this tendency differentiates their song from the other, dissimilarity of authors within same category makes them unrecognizable.

Table 7. Confusion matrix for genre prediction with best classification setup

Predicted **Actual**	**Pop**	**Rock**	**Arabesk**
Pop	280	23	62
Rock	31	209	26
Arabesk	47	5	365

Evaluating the release date prediction performance is not an easy task since the determined boundaries for categorical labeling is intuitive. However, the confusion matrix in Table 8 clearly indicates that most recognizable category is the interval between 1993 and 2006. We can attribute this result to the free-style writing preferences of pop music authors in conjunction with the growth of pop music in popularity in young population in those years. This trend created an unrestricted grammatical structure and simpler vocabulary borrowed from daily conservations.

Table 8. Confusion matrix for release date prediction with best classification setup

Predicted **Actual**	**1972-1993**	**1994-2006**	**2007-2014**
1972-1993	219	65	9
1994-2006	103	410	49
2007-2014	21	71	101

Table 9 lists top features selected by chi-square metric. According to the list, the most informative features for all attribute categories are those relevant to line length statistics. Global text statistics such as number of words, number of characters and number of unique words, stand for the second most contributing feature set. Among bag-of-words, the words "aşk" (the act of loving), "sevgi" (the act of liking), "vücut" (body), "kalp" (heart), "barış" (peace), "ümit" (hope), "bir" (a or one) and "yıl" (year) are the most representative. It seems that, in spite of their wide use, the frequencies of some common words such as "aşk", "sevgi" and "ümit" are effective in determining the writing preferences. Most informative character N-grams are "lar", "ik", "dim", "dü", "miş" and "da", which refer to suffix terms for making plural, past tense for first person plural, past tense for first person singular, past tense for third person singular, past perfect tense for third person singular, and stating location (e.g. "at", "in" or "on" in English) respectively. This indicates that character N-grams can essentially capture the distribution of certain suffixes, thus, determine the grammatical preferences in writing.

Table 9. Most descriptive features for each category

Author	Genre	Release Date
Difference between longest and shortest line lengths	Number of characters	Difference between longest and shortest line lengths
Number of words	Difference between longest and shortest line lengths	Average line length
Number of characters	Number of unique words	Word frequency of "Aşk"
Number of unique words	Word frequency of "Bir"	Average line length
Word frequency of "Barış"	N-gram frequency of "he"	Word frequency of "Sevgi"
Word frequency of "Bir"	N-gram frequency of "da"	N-gram frequency of "urb"
Average line length	Average line length	Word frequency of "Kalp"
Word frequency of "Vücut"	N-gram frequency of "nda"	Word frequency of "Ümit"
N-gram frequency of "lar"	Word frequency of "Gönül	N-gram frequency of "ik"
Word frequency of "Aşk"	N-gram frequency of "lar"	N-gram frequency of "dim"
N-gram frequency of "her"	N-gram frequency of "dü"	Word frequency of "Barış"
N-gram frequency of "da"	N-gram frequency of "miş"	Word frequency of "Yıl"

4 Conclusions

Analyzing lyrics data by quantitative techniques is a promising task to provide an effective means of organizing, accessing and manipulating music content in digital libraries. In this article, we present a comprehensive study to infer valuable knowledge from lyrics content using intelligent data analysis techniques. The contribution of the study is fourfold.

First, a complete framework is introduced for music libraries to estimate three different music meta-data attributes, i.e. author, genre, and release date, using only song lyrics content. To the best of our knowledge, this is the first attempt for predicting author and release date of a song from its lyrics.

Second, a number of relevant feature sets are evaluated and incorporated into a single feature encoding scheme to represent song lyrics. The results have shown that some particular features proposed in this study, such as the difference between the longest and shortest line lengths can help in predicting certain attributes. This new representation scheme can be used in several applications such as music classification, information retrieval and recommendation systems.

Third contribution is the large data set containing carefully annotated Turkish song lyrics, which is gathered and made publicly available in this study. The dataset is public accessible at www.baskent.edu.tr/~hogul/lyrics. We anticipate that researchers in information retrieval, NLP and music analysis communities will benefit from this data set in their future experimental studies.

Finally, lyrics classification task in general is addressed for the first time in the context of Turkish language. We have shown that incorporation of some language-specific pre-processing steps, such as stemming, can improve the classification and retrieval ability. Experimental results suggest the use of proposed models as complementary tools in music information retrieval applications not only for Turkish lyrics but also for songs from other languages.

References

1. Adriano, A., Rodrigo, M., Ribeiro, R.P., Silla J.C.: Nordic Music Genre Classification Using Song Lyrics, 19th Int Conf on App. Natural Language to Information Systems (2014).
2. Akın, M.D., Akın, A.A.:Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi: ZEMBEREK. Elektrik Mühendisliği 431, 38 (2007).
3. Angi, C.: Müzik Kavramı ve Türkiye'de Dinlenen Bazı Müzik Türleri. Idil 2, 59-81 (2013).
4. Cavnar, W.B., Trenkle J.M.: N-gram-based text categorization. 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175 (1994).
5. Costa, C.H.L., Valle, J.D., Koerich A.L.: Automatic classification of audio data. IEEE International Conference on Systems, Man and Cybernetics (2004).
6. Debaecker, J., Widad, M.H.: Music indexing and retrieval: evaluating the social production of music metadata and its use. Facets of Knowledge Organization, 353-363 (2011).
7. Fell, M., Sporleder, C.: Lyrics-based Analysis and Classification of Music. COLING 2014, 620-631 (2014).
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11, 10-18 (2009).
9. Iliev, R., Dehghani, M., Sagi, E.: Automated text analysis in psychology: methods, applications, and future developments, Language and Cognition 7, 265-290 (2014).
10. Kibriya, A.M., Frank, E.., Pfahringer, B.., Holmes, G.: Multinomial naive Bayes for text categorization revisited. Australian Conf on Artificial Intelligence. 488–499 (2004).
11. Kırmacı. B., Oğul, B.: Evaluating text features for lyrics-based songwriter prediction, IEEE 19th International Conference on Intelligent Engineering Systems (2015).
12. Mayer, R., Neumayer, R., Rauber, A.: Rhyme and style features for musical genre classification by song lyrics. 9th International Conference on Music Information Retrieval, pp. 337–342 (2008).
13. Moutselakis, E.V., Karakos, A.S.: Semantic Web Multimedia Metadata Retrieval: A Music Approach. 13th Panhellenic Conference on Informatics (2009).
14. Orio, N.: Music retrieval: a tutorial and review. Foundations and Trends in Information Retrieval 1, 1–90 (2006).
15. Sarkar, R., Saha, S.K.: Music genre classification using EMD and pitch based feature, Eighth International Conference on Advances in Pattern Recognition (2015).
16. Schedl, M., Gómez, E., Urbano, J.: Music Information Retrieval: Recent Developments and Applications. Foundations and Trends in Information Retrieval 8, 127-261 (2014).
17. van Zaannen, M., Kanters, P.: Automatic Mood Classification Using TF*IDF Based on Lyrics. 11th International Society for Music Information Retrieval Conference (2010).
18. Xia, Y., Wang, L., Wong, K.F., Xu, M.: Sentiment vector space model for lyric-based song sentiment classification. 46th Annual Meeting ACL HLT, 133-136 (2008).