Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 21, Sayı 2, 322-331, 2017

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 21, Issue 2, 322-331, 2017

DOI: 10.19113/sdufbed.88303

Turkish Music Genre Classification using Audio and Lyrics Features

Önder ÇOBAN *[1]

[1]Cukurova University, Faculty of Engineering Architecture, Department of Computer Engineering, 01330, Adana

(Alınış / Received: 30.12.2016, Kabul / Accepted: 17.04.2017, Online Yayınlanma / Published Online: 06.05.2017)

Keywords
Music genre classification,
Lyrics analysis,
Word2vec,
Audio classification,
Machine learning

Abstract: Music Information Retrieval (MIR) has become a popular research area in recent years. In this context, researchers have developed music information systems to find solutions for such major problems as automatic playlist creation, hit song detection, and music genre or mood classification. Meta-data information, lyrics, or melodic content of music are used as feature resource in previous works. However, lyrics do not often used in MIR systems and the number of works in this field is not enough especially for Turkish. In this paper, firstly, we have extended our previously created Turkish MIR (TMIR) dataset, which comprises of Turkish lyrics, by including the audio file of each song. Secondly, we have investigated the effect of using audio and textual features together or separately on automatic Music Genre Classification (MGC). We have extracted textual features from lyrics using different feature extraction models such as word2vec and traditional Bag of Words. We have conducted our experiments on Support Vector Machine (SVM) algorithm and analysed the impact of feature selection and different feature groups on MGC. We have considered lyrics based MGC as a text classification task and also investigated the effect of term weighting method. Experimental results show that textual features can also be effective as well as audio features for Turkish MGC, especially when a supervised term weighting method is employed. We have achieved the highest success rate as 99,12% by using both audio and textual features together.

Şarkı Sözü ve Ses Niteliklerini Kullanarak Türkçe Müzik Türü Sınıflandırması

Anahtar Kelimeler
Müzik türü sınıflandırması,
Şarkı sözü analizi,
Word2vec,
Ses sınıflandırması,
Makine öğrenmesi

Özet: Müzik Bilgi Getirimi (MIR) son yıllarda popüler bir araştırma alanı olmuştur. Bu bağlamda, araştırmacılar müzik türü, sevilen şarkıların tespiti ve otomatik çalma listesi oluşturma gibi önemli problemlere çözüm üretmek için müzik bilgi sistemleri geliştirmişlerdir. Önceki çalışmalarda üst-veri bilgisi, şarkı sözleri ya da müziğin melodik içeriği nitelik kaynağı olarak kullanılmıştır. Ancak, şarkı sözleri genellikle MIR sistemlerinde kullanılmamış ve özellikle Türkçe için bu alanda yapılan çalışma sayısı yetersiz kalmıştır. Bu çalışmada, ilk olarak, her bir şarkıya ait ses dosyası eklenerek daha önce oluşturduğumuz Türkçe şarkı sözlerinden oluşan Türkçe MIR (TMIR) veri kümesi genişletilmiştir. İkinci olarak, ses ve metinsel niteliklerin birlikte ve ayrı kullanıldıklarında Müzik Türü Sınıflandırması (MGC) üzerindeki etkisi incelenmiştir. Metinsel nitelikler word2vec ve kelime torbası gibi nitelik çıkarım modelleri ile şarkı sözlerinden çıkarılmıştır. Deneyler Destek Vektör Makinesi (SVM) algoritması ile gerçekleştirilmiş ve nitelik seçimi ile farklı nitelik gruplarının MGC üzerindeki etkisi incelenmiştir. Şarkı sözü tabanlı MGC bir metin sınıflandırma işlemi olarak ele alınmış, ayrıca terim ağırlıklandırma yönteminin etkisi incelenmiştir. Deneysel sonuçlar ses niteliklerinin yanı sıra özellikle denetimli bir ağırlıklandırma yöntemi kullanıldığında metinsel niteliklerin de MGC için etkili olabileceğini göstermiştir. Metinsel nitelikler ses nitelikleri ile birlikte kullanılarak en yüksek %99,12 oranında başarı elde edilmiştir.

1. Introduction

As the number of digital and online music increases, organizing musics has emerged as a major problem to be solved. Users need easy and fast access to related music using its diverse attributes such as genre, style and year. The artist, user tags and music genre or mood classification are remarkable research areas in automatic music classification which has an important place in music information retrieval (MIR) [1]. Developed music

* *Corresponding author: ocoban@cu.edu.tr*

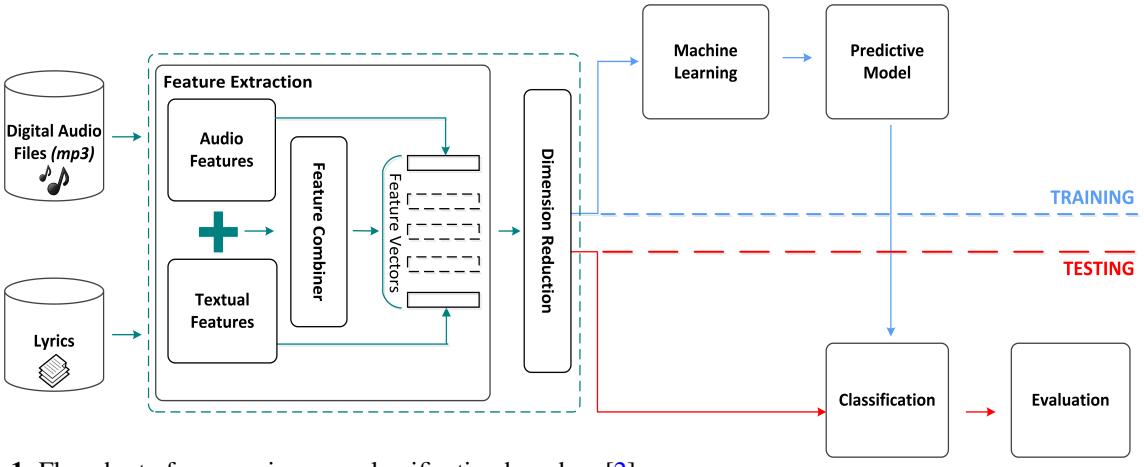

Figure 1. Flowchart of our music genre classification based on [2].

classification systems are usully based on melodic content. However, an opinion has gained importance recently which claims that lyrics can also be used alone or together with melodic content. In this context, lyrics are used for the purpose of improving the system performance in Music Genre Classification (MGC) field [3, 4]. Therefore, we have investigated the effect of lyrics on Turkish MGC in our previous work [5]. Our results show that the lyrics may be useful in Turkish MGC. When we examine the literature, we have seen that there is not enough study, which has directly focused on MGC of songs, especially for Turkish. In the previous studies, researchers generally focus on rhythmic similarity detection [6] and makam classification [7, 8] of Turkish songs. However, to our best knowledge there is no previous work which employs audio and lyrics features for automatic MGC of Turkish songs, although there are a few lyrics based works for Turkish [5, 9]. Therefore, in this study, we have extended our previously created Turkish Music Information Retrieval (TMIR) dataset by including the audio file of each song. Then, we have performed the MGC on TMIR dataset using both lyrics and audio signals. Our main goal is to investigate whether lyrics or audio signal of music is more effective for Turkish MGC. We also try to answer the question that, can the MIR performance be improved when audio and lyrics features are utilized together. For these purposes, we have extracted textual and audio features from lyrics and MP3 formatted audio signals respectively. We have extracted textual features by using different representation models such as traditional Bag of Words (BoW) and NGram. We have considered lyrics based MGC as a classical text classification task [10] and employed different term weighting methods. We have also extracted the most commonly used timbral features from audio signals such as RMS and MFCC's [11].

We have the following contributions in this paper: (a) we have created the TMIR dataset which can be used in Turkish MGC studies; (b) we have performed MGC of Turkish songs using both audio and lyrics features; (c) we have employed Word2Vec (W2V) model and an improved term weighting scheme while representing the lyrics features; (d) we have also investigated the effect of feature selection and feature combination on results.

The rest of the paper is organized as follows: The related works are summarized in Section 2. Methods for extracting both lyrics and audio features, feature combination, dimension reduction, and classification are described in Section 3. The detailed information about TMIR dataset is given in Section 4. We show our experimental results in Section 5 and analyze the best feature group combination. In the final section, we summarize results and give our suggestions.

2. Related Works

In MIR process, lyrics are used as feature source alone or together with its melodic content [4]. In this context, there are previous studies which use lyrics and audio signals both separately or together for automatic MGC. Yang and Lee have used lyrics with melodic content and they have detected mood from lyrics [12]. Mayer and Rauber have analyzed lyrics, which are collected from web, semantically and structurally and they classified lyrics as tematic aspect [13]. Van Zaanen and Kanters employed TF-IDF weighting to statistical features which are extracted from lyrics and they employed these features in mood detection [14]. McKay et al, investigated the effect of lyrics on music genre classification using statistical and structural text features [1]. Mayer et al, utilized the lyrics in music genre classification. They have also analyzed the impact of feature groups including BoW, statistical, part of speech, and rhyme pattern features on classification results [10]. Kirmaci and Ogul have automatically detected music genre and songwriter by utilizing four different feature sets, which are obtained from Turkish lyrics [9]. Ogul and Kırmacı have also used lyrics to automatically detect the meta-data of a Turkish song. They have employed different feature sets extracted from lyrics and focused on attributing the author, genre and release date of songs [15]. Coban and Ozyer performed automatic MGC on Turkish lyrics. They have performed a comparative analysis where the feature set, term weighting method and classifier are different [5]. Yaslan and Cataltepe have performed MGC using audio features and employed different classifiers and feature selection methods [16]. Dhanaraj and Logan have

Table 1. The summary of applied preprocessing steps on lyrics for different feature extraction models

Feature Model	Preprocessing Steps						
	Lowercase Conversion	Removing Punctuations	Stemming	Removing Stopwords	Text Normalization	ASCII Conversion	Term Weighting
BoW	+	+	+	+	+	-	+
NGram	+	+	-	-	+	-	+
SSTF	+	-	-	-	-	-	-
W2V	+	+	+	+	+	+	-

used lyrics and audio features for automatic prediction of hit songs [17]. Cataltepe et al, carried out MGC using MIDI and audio features such as timbral, rhythmic, and pitch content features [18]. McKay and Fujinaga combined audio (MP3), cultural (metadata) and symbolic (MIDI) features and analyzed their effect on results [19]. However, we have not found any previous research work that performs MGC on Turkish songs using both audio and lyrics features. In this context, this study is quite important and valuable for Turkish MGC.

3. Methods

We describe our methodology under five sub-headings, and the flowchart of our system which is given in Figure 1. In our flowchart, we extract audio and textual features from audio files and lyrics respectively. Then, we apply dimension reduction and train a machine learning algorithm. In final step, we employ our predictive model to classify each test sample. In this section, first, we clarify our employed feature extraction steps to obtain both textual and timbral audio features. Then, we describe feature combination, dimension reduction, classification, and evaluation stages respectively.

3.1. Extracting audio features

Music is commonly analyzed on audio or symbolic level in MIR systems. Songs are represented by low level features which are calculated by using sound waves or records. Therefore, we have extracted timbral audio features using jAudio[1] library which is one of the commonly used tools in MIR studies. The jAudio library has been developed to calculate features from audio signals in formats such as MP3, AIFF, and WAV. It is Java based and enables to use some other abilities such as handling dependencies, support for multidimensional features, and extensibility. For these reasons, we have extracted 27 distinct features implemented in jAudio [20]. These features may be one dimensional (e.g., RMS) or multidimensional (e.g., MFCC's) [21]. A list which includes a subset of features is given in Table 2.

3.2. Extracting textual features

Except for instrumental songs, lyrics can be considered as a textual content and by processing this content with classical text processing techniques MIR can be achieved [5, 9, 10]. Therefore, in this phase, we have considered the MGC as a classical text classification task

Table 2. A non-exhaustive list of jAudio features

Feature	Description
Zero Crossing	Calculated using # of times that the singal crosses zero
RMS	A measure of the audio signal power
Compactness	Provides an indication of the noisiness of the singal
MFCC	Coefficients derived from a cosine transform of the real logarithm of the short-term power spectrum of a sound [11]
Beat Histogram	Constructs a histogram representing the rhythmic regularities

and extracted textual features from lyrics. For this purpose, we have employed the following steps for lyrics text processing.

3.2.1. Preprocessing

The lyrics are textual content which has their own structural attributes and comprised of chorus, verse or segment blocks. In addition, the lyrics are often contain rhymes and arranged as adhere to a specific template. In MGC, such statistical and structural features are easier to obtain from lyrics compared to the melodic content. However, lyrics need to be passed to some specific preprocessing techniques [4, 22]. Therefore, we have applied following preprocessing steps (see Table 1) on Turkish lyrics contained in our TMIR dataset.

Firstly, we have applied lowercase conversion on lyrics. All characters and non-informative contents which are not considered as feature are removed from content. A single blank line between segments is also provided and choruses are not reduced to a single segment. After this, we have used three feature extraction models including BoW, NGram, SSTF (Structural and Statistical Text Features). In addition, we have represented the lyrics using the W2V word embedding method, which is recently quite popular in other fields such as information retrieval, text classification and sentiment analysis [23]. In this context, we have applied different preprocessing steps according to the feature model. In BoW and NGram models, we have removed punctuations and characters except for letters from the text. However, we have not applied term frequency and length filters in these models. In all four models, we have employed Zemberek[2], which is an open source natural

[1] http://jmir.sourceforge.net/jAudio.html

[2] https://code.google.com/p/zemberek/

Table 3. The features contained in SSTFI and SSTFII groups

	Features			
SSTFI	*Words Per Sentence Average*	*Stopwords Ratio*	*Average Length Per Line*	*Lines Per Segment Average*
	Number of Rhymes - (ABAB | BABA)	*Number of Lines*	*Average Length Per Word*	*Number of Rhymes - (AA | BB)*
	Average Line Per Segment	*Number of Empty Lines*	*Average Length Per Sentence*	*Number of Rhymes - (ABBA | BAAB)*
	Number of Rhymes - (AABB | BBAA)	*Number of Segments*	*Number of Sentences*	*Number of Unique Rhyme Words*
SSTFII	*Average Syllable Count Per Word*	*Number of Words*	*Average Word Per Line*	*(-) Hyphen Frequency*
	Average Character Count Per Word	*Number of Numerical Values*	*(") Double Quotation Frequency*	*(...) Ellipsis Frequency*
	Characters Per Word Variance	*Vocabulary Richness*	*(:) Colon Frequency*	*(!) Exclamation Frequency*
	Average Unique Words Per Line	*Number of Unique Words*	*(;) Semicolon Frequency*	*(,) Comma Frequency*
	Words Per Line Variance	*Punctuation Ratios*	*(?) Question Mark Frequency*	*(*) Asteriks Frequency*

language processing toolkit for Turkish, to perform semantic actions [24]. We have utilized Turkish stopwords list (see Figure 2) contained in Lucene[3] API while removing stopwords and calculating the stopwords ratio in textual content. We have also applied text normalization which is mostly applied in sentiment analysis studies to reduce feature space as in [25]. We have performed the ASCII conversion only in W2V model, due to it is not working compatible with Turkish characters [26]. The summary of our preprocessing steps on lyrics for different feature extraction models is given in Table 1.

Figure 2. Turkish stopwords in Lucene API.

3.2.2. Lyrics features

The lyrics features are extracted in four different models including SSTF, BoW, NGram, and W2V.

BoW Features: The sample representation in training data has considerable impact on the ability to make generalization of learning system. In BoW model, this representation is often done by associating word-frequency to convert it suitable for classification [27]. In this way, each word and its frequency represents a feature and its value (TF weight) in the textual content respectively. In this model, word orders in text considered to be not important [28].

NGram Features: NGram model is alternative feature extraction model in text classification. It can be applied on two different ways, including word and character level. But the character level model is generally more successful than word level [29]. In character level ngram model, the features are n-character strings extracted from the textual content. In this aspect, it is independent of language and being strong against the cases such as use of misspelling and abbreviations [30]. In this study, we have used three different character level ngram features including bigram, trigram, and four-gram.

SSTF: This model is commonly employed in such authorship attribution detection, text genre recognition, and text classification [31, 32] fields that those use statistical and structural attributes of texts. In this study, we have obtained 45 unique SSTF from lyrics by taking into account the features used in previous studies [1, 3, 10]. We have divided this features into three different subgroups including SSTFI (structural), SSTFII (statistical) and SSTFIII (POS Tags Frequencies). The features contained in SSTFI[4] and SSTFII are given in Table 3. In POS (Part of Speech Tags) group, we have used frequency of *prepositions*, *verb*, *noun*, *adjective*, *pronoun*, *conjunction*, *verb+adjective*, *noun+verb*, and *noun+adjective* respectively. We have also included rhyme pattern features (SSTFU), which are obtained using method in [10], to SSTFI group. In previous studies, which employed rhyme pattern (AA, AABB, ABAB, ABBA) features, the similarity between last syllables of words is just not sufficient. The sound similarity is also sought to ensure rhyme condition. However, in this study, syllabic similarity is considered sufficient, due to the words are read as written in Turkish. We have considered all rhymes (two or more sound similarity) except for assonance (only sound similarity) in rhyme pattern detection phase. In Turkish, sound similarity is also seen between same words (in Turkish "redif") at the end of the lines. However, we do not check whether such words contain redif or rhyme and we also consider redif as rhyme.

W2V: This model proposed to produce word embeddings by translating words to vectors. The mathematical theory behind the W2V model is described in [33, 34]. The main idea of this model is mapping each word into n dimensional vector and detect the semantic similarity of words by calculating the distance [35]. The W2V model can generate the word vectors using two different model architecture including Continuous Bag of Words (CBoW) and Skip-Gram. The CBoW predicts the current word by using a window of surrounding context words, in

[3] http://lucene.apache.org/

[4] Includes rhyme pattern features (AA, AABB, ABAB, ABBA)

Table 4. Distribution of artists and number of lyrics and audio samples among categories in TMIR data

Category	Artists					Total # of lyrics	Total # of audio
Pop	*Gülşen*	*Ajda Pekkan*	*Murat Boz*	*Demet Akalın*	*Gökhan Özen*	*250*	*250*
Rap	*Ceza*	*Kolera*	*Sagopa Kajmer*	*Killa Hakan*	*Allame*	*250*	*250*
Rock	*Emre Aydın*	*Feridun Düzağaç*	*Haluk Levent*	*Cem Karaca*	*Şebnem Ferah*	*250*	*250*
Folk Music	*Güler Duman*	*Hüseyin Turan*	*Yavuz Bingöl*	*Musa Eroğlu*	*Belkıs Akkale*	*250*	*250*
Arabesque	*Müslüm Gürses*	*Orhan Gencebay*	*Ferdi Tayfur*	*Hakan Taşıyan*	*İbrahim Tatlıses*	*250*	*250*
Total						**1250**	**1250**

contrary to the Skip-Gram which uses current word to predict surrounding window of context words. The order of context words is not important in CBoW which is faster than the Skip-Gram. However, the Skip-Gram outperforms CBoW when used for infrequent words[5]. In this work, we have represented lyrics using W2V embeddings as in [23].

After the feature extraction, we have represented each sample in dataset using the Vector Space Model [36] which is commonly used model in information retrieval.

3.2.3. Term weighting

Term (feature) weighting process plays a significant role in text classification and assigns appropriate weighting values to the terms in sample vector. The assigned weight helps to increase discriminating power of the related term in sample. In this study, term weighting is just applied on BoW and NGram features by using Improved Term Weighting (ITW) method [37] which outperforms traditional weighting methods such as TF and TF-IDF [38]. The mathematical representation of the supervised ITW method is given in following equation:

$$W_{ITW}(i,k) = tf_{ik} * log(\frac{N}{n_k} + 0{,}01) * \frac{ttn_k}{tc_k * |C_i|} \quad (1)$$

where W denote weighting metod, while tf_{ik}, N, n_k, ttn_k, tc_k, and $|C_i|$ represent the observed frequency of t_k in lyric L_i, total number of lyrics, number of lyrics in which t_k occurs, number of lyrics in related category which contain t_k, number of categories in which t_k occurs, and the total number of lyrics in related category respectively.

3.3. Feature combination

We have also combined four different feature groups extracted from lyrics (BoW, Nram, SSTF, and W2V) and audio features to investigate the effect on results. In this phase, while combining the feature groups we have not applied re-weighting on feature values (or calculated weights) in each model.

3.4. Dimension reduction

In this phase, we have performed dimension reduction by using correlation based feature selection (FS) algorithm (CFS) which is proposed for machine learning tasks [39]. We prefer the CFS algorithm, due to it automatically decides how many features worth to be selected. CFS ranks feature subsets according to a correlation based heuristic evaluation function. The bias of this function is toward subsets which contain highly correlated features in related class and uncorrelated with others. In other words, the acceptance of a feature depends on the extent to which it predicts classes in areas of the sample space not already predicted by other features. In this phase, we have employed CFS algorithm using two different search methods including Best First Search (BF) and Genetic Search (GS) to investigate the effect on results.

[5] https://code.google.com/archive/p/word2vec/

Table 5. The feature groups and codes used in experiments

Feature Group / Model			Code
Lyrics	SSTF	Structural Features (see Table 3)	SSTFI
		Statistical Features (see Table 3)	SSTFII
		Part of Speech Frequencies (POS Tags)	SSTFIII
		Rhyme Patterns (AA, AABB, ABAB, ABBA)	SSTFU
		SSTFI + SSTFII + SSTFIII	SSTF
	BoW	BoW + Stemming	BoWS
		BoW + Removing Stopwords	BoWRS
		BoW + Stemming + Removing Stopwords	BoW
	NGram	Bigram	NGB
		Trigram	NGT
		Four-gram	NGF
	W2V	CBoW	WVC
		Skip-Gram	WVS
Audio		Audio Features (see Table 2)	AUD
Combined		Lyrics + Audio Features (see Table 9)	CMB

3.5. Classification and evaluation

The classification is assigning most suitable category from training data categories to previously unseen sample. In this study, the Support Vector Machine (SVM) algorithm, which is most common and usually the most successful method in MIR field, employed as a classifier [40]. While evaluating the performance of classifier, we have employed the accuracy (Acc) measure which is one of the most used metrics in machine learning [41]. Let TP, FP, TN, and FN represent number of true positives, false positives, true negatives, and false negatives respectively. Then, the Acc measure can be defined as follows:

$$Acc = (TP + TN)/N \quad (2)$$

where $N = TP + FP + TN + FN$ which equals to the total number of samples in test data.

4. Dataset

There does not exist a publicly available dataset for Turkish MIR and MGC. Therefore, we have created TMIR dataset which comprises of both lyrics and audio files of Turkish songs. While creating the TMIR dataset, we have collected

Table 6. The number of unique features in different feature groups before and after the feature selection

FS	Feature Groups													
	BoWS	BoWRS	BoW	NGB	NGT	NGF	SSTFI	SSTFII	SSTFIII	SSTFU	SSTF	WVC	WVS	AUD
No FS	5238	33043	5144	795	7155	32273	16	20	9	5	45	500	500	2332
CFS_{BF}	194	84	181	325	661	618	8	3	7	3	17	429	105	15
CFS_{GS}	1307	10367	579	202	2179	13947	8	9	7	3	23	461	242	851

Table 7. Classification accuracies without feature selection for different feature groups and weighting strategies (%)

Weighting Scheme	Feature Groups													
	Weighting Applied						No Weighting							
	BoWS	BoWRS	BoW	NGB	NGT	NGF	SSTFI	SSTFII	SSTFIII	SSTFU	SSTF	WVC	WVS	AUD
W_{TF}	67,52	67,52	66,72	63,52	67,12	70,32	**55,52**	**52,40**	**42,28**	**34,72**	**63,36**	**61,04**	**52,00**	**97,68**
$W_{TF\text{-}IDF}$	67,68	67,52	66,72	61,04	66,72	70,32								
W_{ITW}	**87,28**	**87,04**	**85,68**	**85,84**	**94,56**	**96,88**								

lyrics and audio files of related songs from a web site[6] and YouTube[7] respectively. After this, we have converted the audio files to MP3 format with a 48 kHz sample rate, 320 kbps bit rate, and 16 bit sample size. Our dataset has songs under the five different categories (music genres) including rock, rap, arabesque, pop, and folk music. In TMIR dataset, musics are selected from five different popular artists which sang the songs in related genre. While deciding to in which genre an artist will be included, we just considered the condition that the artist signs songs only in related music genre. Then, we have randomly selected 50 songs that each artist sings and created the TMIR dataset by extending our previous dataset which includes automatically collected Turkish lyrics [5]. In this phase, we have also detected a few lyrics which are contained in different genres at the same time. Therefore, we have removed such lyrics and related audio signals from the dataset. As can be seen from the Table 4, the TMIR dataset contains equally distributed (balanced) lyrics and audio files of 2500 Turkish songs at total.

5. Results

In this section, we present and analyze our results. First, we clarify the configuration and then, present our experimental results.

5.1. Configuration

We have utilized the SVM algorithm using the LIBSVM[8] package and tuned with default parameters and Linear kernel. We have used the WEKA [42] implementations of the CFS, BF and GS algorithms with default parameters. We have also set the dimension of word vectors[9] as 500 in W2V model and used with default parameters. To validate our predictive model, we have used n-fold cross-validation (CV) technique which is most commonly used method in machine learning [41]. This method randomly divides dataset n times into n complementary subsets. It considers one of these subsets as test set while taking the rest as training set. The total error of the model would be mean error of error rates obtained from each of the n iterations [43]. In this study, we have configured the CV with 10 folds.

5.2. Experimental results

In this section, we present our experimental results. To obtain the results, we have processed both audio signal and lyrics of each related song in TMIR dataset. For this purpose, we have employed the techniques which described in Section 3. After the processing phase, we have extracted lyrics and audio features. We have obtained lyrics features using four different models including BoW, NGram, SSTF, and W2V. We have divided these lyrics features into subgroups to investigate the effect on results. A total of 45 features contained in SSTF model are also divided into three groups. In BoW model, the features are obtained in three different ways in order to analyze the effects of stemming and removing stopwords. We have employed NGram features on different character levels including bigram, trigram, and four-gram. In addition, we have employed the W2V model using both CBoW and Skip-Gram architectures. We take the most successful feature group from the both lyrics and audio features and combine them with each other to investigate effect on results. While combining the feature groups, we do not apply re-weighting on feature values (or calculated weights). All these feature groups and codes used in experiments are given in Table 5. In addition, we have employed traditional TF and TF-IDF weightings on BoW and Ngram features to compare with ITW method. After the feature extraction, we have reduced the feature space using the CFS feature selection algorithm. We have employed the CFS algorithm using both BF and GS search methods to investigate the effect on results.

The numbers of unique features in different feature groups before and after the feature selection are given in Table 6. As can be seen from the table, BoW and NGram models have high dimensionality of feature space which is major problem in classical text classification. The audio features has also higher feature space when compared with both W2V and SSTF. In audio feature groups, we have obtained 2332 unique features, although the jAudio has 27 distinct features. The reason for this that some of these audio features are one dimensional or multidimensional. In addition, the jAudio automatically produces new features

[6] http://sarki.alternatifim.com/
[7] https://www.youtube.com/
[8] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[9] http://deeplearning4j.org/word2vec

Table 8. Classification accuracies after the feature selection for audio and lyrics features (%)

FS	Feature Groups													
	BoWS	BoWRS	BoW	NGB	NGT	NGF	SSTFI	SSTFII	SSTFIII	SSTFU	SSTF	WVC	WVS	AUD
$CFS_{\mathbf{BF}}$	**83,20**	65,20	**81,84**	**84,80**	88,48	89,44	55,76	47,20	45,92	32,88	57,12	60,08	45,04	96,32
$CFS_{\mathbf{GS}}$	75,60	**71,76**	68,00	76,24	**88,80**	**94,32**	**55,76**	**51,04**	**45,92**	**32,88**	**62,88**	**61,92**	**47,04**	**98,00**

Table 9. Classification accuracies for combined feature groups

Features	Acc (%)	Features	Acc (%)
$CMB_{\mathbf{SSTF + BoWS}}$	83,28 ↑	$CMB_{\mathbf{NGF + WVC}}$	88,96 ↓
$CMB_{\mathbf{BoWS + NGF}}$	94,80 ↑	$CMB_{\mathbf{AUD + WVC}}$	95,68 ↓
$CMB_{\mathbf{NGF + SSTF}}$	94,64 ↑	$CMB_{\mathbf{SSTF + AUD}}$	98,00 ↔
$CMB_{\mathbf{BoWS + WVC}}$	68,96 ↓	$CMB_{\mathbf{AUD + BoWS}}$	**99,12** ↑
$CMB_{\mathbf{WVC + SSTF}}$	67,12 ↑	$CMB_{\mathbf{NGF + AUD}}$	98,72 ↑

Table 10. The obtained results for different measures

Feature Set	Measure				
	FPR	TPR	RMSE	PC	PI
AUD	0,01	0,98	0,08	98,00	2,00
SSTF	0,10	0,48	0,31	62,88	37,12
NGF	0,02	0,90	0,13	94,32	5,68
W2V	0,10	0,51	0,33	61,92	38,08
BoWS	0,05	0,74	0,22	83,20	16,80
$CMB_{AUD+BoWS}$	0,00	1,00	0,06	99,12	0,88
$CMB_{SSTF+BoWS}$	0,05	0,75	0,22	83,24	16,76
$CMB_{BoWS+NGF}$	0,02	0,91	0,13	94,80	5,20
$CMB_{NGF+AUD}$	0,01	0,97	0,07	98,72	1,28
$CMB_{NGF+SSTF}$	0,02	0,90	0,13	94,64	5,36
$CMB_{SSTF+AUD}$	0,01	0,98	0,08	98,00	2,00
$CMB_{W2V+SSTF}$	0,09	0,58	0,30	67,12	32,88

using metafeatures from existing features [21]. We have obtained our experimental results in two different ways including before (No FS) and after ($CFS_{\mathbf{BF}}$, $CFS_{\mathbf{GS}}$) the feature selection to investigate the effect of feature selecion process. First, we have employed both lyrics and audio features without feature selection and obtained the results which are given in Table 7. As can be seen from the table, we have also used different weighting methods on BoW and NGram features to compare with ITW. After this step, we have applied feature selection for all feature groups and obtained the results which are given in Table 8. Finally, we have selected the most successful sub-feature groups and combined with each other to investigate the effect on results. For example, we take the most successful subgroups which are the BoWS features selected with $CFS_{\mathbf{BF}}$ and NGF features selected with $CFS_{\mathbf{GS}}$ respectively. Then, we combine these two groups with each other and rename as $CMB_{\mathbf{BoWS + NGF}}$. These combined feature groups which we employed in the experiments are given in Table 9.

When we consider the Table 7, our experimental results show that NGF features outperform other lyrics features in all cases. In NGram model, the classification success is directly proportional to character level. BoW features are more successful than the NGB and NGT groups. In addition, applying both stemming and removing stopwords together has decreased classification success in BoW model. The ITW method outperforms traditional weighting methods. In this aspect, the ITW has significant effect on performance of both BoW and NGram features. Therefore, we have decided to use ITW weighted BoW and NGram features in the rest of the experiments. The SSTFU group has also lowest success in all lyrics features. The CBoW architecture is more successful than the Skip-Gram in W2V model. In SSTF model, we have obtained best results when we use all sub-feature groups. An important point to mention here is that the audio features are more successful than lyrics features in our MGC. As can be seen from the table, the AUD feature group has the highest success (97,68) among all feature groups. When we consider the Table 8, experimental results show that CFS generally decreased the classification success, especially when employed with BF search. Therefore, the GS is generally more effective than BF. In addition, the CFS has increased the success (98,00) of audio features when used with GS. However, the GS generally selects more feature (see Table 6) than BF, especially on large feature space. This situation generally has positive impact on results but also has led to a disadvantage in terms of performance.

In the final phase of the our experiments, we have combined the most successful (after feature selection) feature groups and obtained 10 different feature set (see Table 9). The feature combination process generally increased the classification success. We have achieved the success rate of 99,12 percent using $CMB_{\mathbf{AUD + BoWS}}$ feature set. This value is the our highest success rate in all experiments. In addition, according to the results in Table 9, we have performed tests for some measures to justify that combined features provide better performance. In this phase, we have obtained the test results for SVM on multiple datasets using different measures including true positive rate (TPR), false positive rate (FPR), root mean squared error (RMSE), percent correct (PC), and percent incorrect (PI). Table 10 presents the results which also show that combined features provide better performance.

6. Discussion and Conclusion

In this study, we have performed automatic MGC on Turkish songs. The main goal of our work is to investigate whether lyrics or audio signal of music is more effective for Turkish MGC. In addition, to answer the question that if the MIR performance can be improved when audio and lyrics features utilized together. For these purposes, firstly, we have created TMIR dataset by including the MP3 formatted audio signals to our previously created lyrics dataset. Secondly, we have employed both audio and lyrics features in MGC process using machine learning techniques. According to the our experimental results, we conclude that lyrics features may be more effective for MGC, especially when supervised term weighting approaches employed such as ITW. The BoW and NGram features are most successful groups in lyrics features,

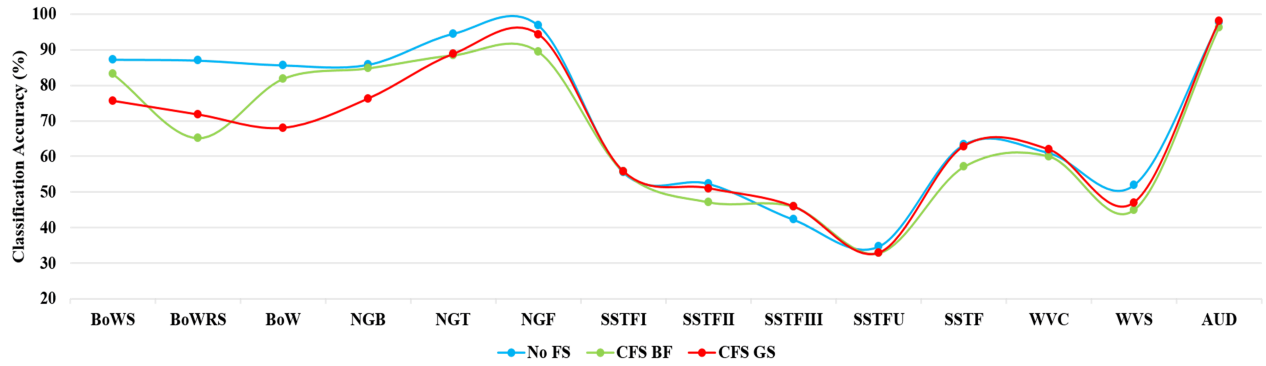

Figure 3. The effect of feature selection on classification accuracy using different search methods for CFS

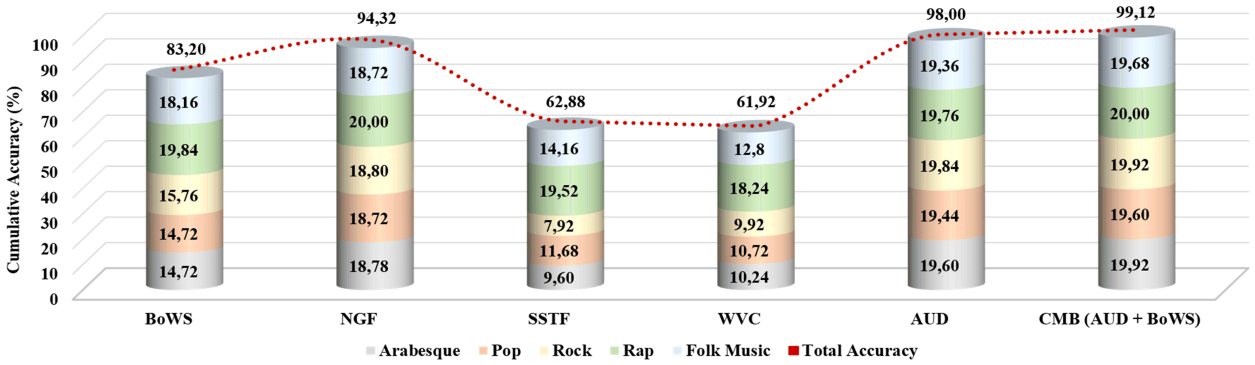

Figure 4. The distribution of cumulative accuracy among music genres for most successful feature groups after the feature selection

however, these models have a disadvantage in terms of high dimensionality of feature space without feature selection. The SSTF and W2V groups are not as successful as BoW and NGram. In SSTF, the majority of the features are not distinctive and this prevents the acquisition of high achievement. The most important reason for this situation is that lyrics are genereally stored without being tied to a specific format and may vary in the digital environment. In addition, the success of the W2V model is not so bad when compared with TF or TF-IDF weighted BoW and NGram models. When considered from this perspective, the W2V model has shown promising performance. Furthermore, the number of features obtained in SSTF and W2V models are being too little compared to the other two models and provide huge performance advantage. The audio features outperform lyrics features in all expriments. This proves the importance of the audio features in MGC field. However, the best performance is obtained using both lyrics and audio features combination. Therefore, we suggest the use of lyrics features as well as audio features.

In addition to the above conclusions, we have shown the effect of feature selection on performance of feature groups in Figure 3. We have also investigated the distribution of cumulative success among music genres. This distribution for the most successful feature groups is given in Figure 4. When we consider Figure 3 and Figure 4, we conclude that the feature selection generally decreases the classification success. The CFS algorithm is generally more effective on high feature space when the BF search employed. Contrary to this, the GS which employed in CFS algorithm is generally more effective on low feature space. In addition, we have investigated selected audio features with $CFS_{\mathbf{BF}}$ and $CFS_{\mathbf{GS}}$. We have observed that MFCC, LPC, Power Spectrum, Magnitude Spectrum, and Method of Moments are most distinctive audio features. This means that audio signals in our dataset have quite strong different frequency components and we have good resolution in low frequencies. When we examine the cumulative distribution of accuracy among genres, we observe that lyrics features are more distinctive in rap and folk music genres respectively. We think the reason for this is that, in rap genre, the lyrics have rich text content and have typical terms and feelings such as sadness and mutiny. The folk music also bears traces of Turkish culture and this makes it more distinctive. However, the audio features are more distinctive in rock and rap genres respectively. This is because the rock and rap songs could be more distinctive for some attributes such as rhythm, beat, tempo, and pitch frequency. These attributes provides obtaining high distinctive audio features in related genres. In addition, the combined features are more distinctive in rap, rock, and arabesque respectively. Therefore, we also conclude that generally the most distinctive genre is rap in Turkish MGC.

In future works, we will investigate the effect of different sample rate and size (or bit depth) on MGC. In this study, we employ only timbral features but we will employ other audio features which can be extracted from both rhythmic and pitch contents. We are also planning to further improve our basic rhyme pattern detection algorithm to detect all Turkish rhymes. Furthermore, we will include Turkish classical music and hymn genres to our TMIR dataset. We

will also share the dataset by making it available online.

References

[1] McKay, C., Burgoyne, J. A., Hockman, J., Smith, J. B., Vigliensoni, G., Fujinaga, I. 2010. Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features. ISMIR, 9-13 August, Utrecht, 213-218.

[2] Sordo, M. 2012. Semantic annotation of music collections: A computational approach. Universitat Pompeu Fabra, Department of Information and Communication Technologies, Doctoral dissertation, 18p, Barcelona.

[3] Hu, X., Downie, J. S. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In Proceedings of the 10th annual joint conference on Digital libraries, 21-25 June, Gold Coast, QLD, 159-168

[4] Ying, T. C., Doraisamy, S., Abdullah, L. N. 2012. Genre and mood classification using lyric features. In Information Retrieval & Knowledge Management (CAMP), 13-15 March, Kuala Lumpur, 260-263.

[5] Çoban, Ö., Özyer, G. T., 2016. Music genre classification from Turkish lyrics. In 2016 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, Zonguldak, 101-104

[6] Holzapfel, A., Stylianou, Y. 2009. Rhythmic Similarity in Traditional Turkish Music. In ISMIR, 26-30 October, Kobe, 99-104.

[7] Alpkoçak, A., Gedik, A. C. 2006. Classification of Turkish songs according to makams by using n grams. In Proceedings of the 15. Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN), Muğla

[8] Kızrak, M. A., Bayram, K. S., Bolat, B. 2014. Classification of Classic Turkish Music Makams. In Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 23-25 June, Alberobello, 394-397

[9] Kırmacı, B., Oğul, H. 2015. Author recognition from lyrics. Signal Processing and Communications Applications Conference (SIU), 16-19 May, Malatya, 2489-2492.

[10] Mayer, R., Neumayer, R., Rauber, A. 2008. Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In ISMIR, 14-18 September, Philadelphia, 337-342.

[11] Zheng, F., Zhang, G., Song, Z. 2001. Comparison of different implementations of MFCC. Journal of Computer Science and Technology, 16(2001), 582-589.

[12] Yang, D., Lee, W. S. 2009. Music emotion identification from lyrics. In ISM'09, 14-16 December, San Diego, 624-629

[13] Mayer, R., Rauber, A. 2010. Multimodal Aspects of Music Retrieval: Audio, Song Lyrics–and Beyond?. In Advances in Music Information Retrieval, 333-363

[14] Van Zaanen, M., Kanters, P. 2010. Automatic Mood Classification Using TF* IDF Based on Lyrics. In ISMIR, 9-13 August, Utrecht, 75-80.

[15] Oğul, H., Kırmacı, B. 2016. Lyrics Mining for Music Meta-Data Estimation. In IFIP International Conference on Artificial Intelligence Applications and Innovations. Thessaloniki, 16-18 September, 528-539.

[16] Yaslan, Y., Cataltepe, Z. 2006. Audio music genre classification using different classifiers and feature selection methods. In 18th International Conference on Pattern Recognition (ICPR'06), 20-24 August, Hong Kong, 573-576

[17] Dhanaraj, R., Logan, B. 2005. Automatic Prediction of Hit Songs. In ISMIR, 11-15 September, London, 488-491.

[18] Cataltepe, Z., Yaslan, Y., Sonmez, A. 2007. Music genre classification using MIDI and audio features. EURASIP Journal on Advances in Signal Processing, 2007(2007), 1-8.

[19] McKay, C., Fujinaga, I. 2008. Combining Features Extracted from Audio, Symbolic and Cultural Sources. In ISMIR, 14-18 September, Philadelphia, 597-602.

[20] McKay, C., Fujinaga, I., Depalle, P. 2005. jAudio: A feature extraction library. In Proceedings of the International Conference on Music Information Retrieval, 11-15 September, London, 600-3.

[21] McEnnis, D., McKay, C., Fujinaga, I., Depalle, P. 2006. jAudio: Additions and Improvements. In ISMIR, 8-12 October, Victoria, 385-386.

[22] Hu, X., Downie, J. S., Ehmann, A. F. 2009 Lyric text mining in music mood classification. American music, 183(2009), 2-209.

[23] Lin, Y., Lei, H., Wu, J., Li, X. 2015. An Empirical Study on Sentiment Classification of Chinese Review using Word Embedding. PACLIC 29, 30 October- 1 November, Shanghai, 258-266.

[24] Akın, A. A., Akın, M. D. 2007. Zemberek, an open source NLP framework for Turkic languages. Structure, 10(2007), 1-5.

[25] Çoban, Ö., Özyer, G. T., 2016. Sentiment classification for Turkish Twitter feeds using LDA. In 2016 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, Zonguldak, 129-132.

[26] Öztürk, M. B., Can, B. 2016. Clustering word roots syntactically. In 2016 24th Signal Processing and Communication Application Conference (SIU), 16-19 May, Zonguldak, 1461-1464

[27] Joachims, T. 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97), 8-12 July, 143-151.

[28] Lewis, D. D. 1992. An evaluation of phrasal and clustered representations on a text categorization task.

In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 21-24 June, Copenhagen, 37-50.

[29] Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E. 2006. Words Vs Characters N-Grams for Anti-Spam Filtering. International Journal on Artificial Intelligence Tools, world Scientific, X(2006), 1-20.

[30] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C. 2002. Text classification using string kernels. Journal of Machine Learning Research, 2(2002), 419-444.

[31] Stamatatos, E., Fakotakis, N., Kokkinakis, G. 2000. Automatic text categorization in terms of genre and author. Computational linguistics, 26(2000), 471-495.

[32] Mayer, R., Neumayer, R., Rauber, A. 2008. Combination of audio and lyrics features for genre classification in digital audio collections. In Proceedings of the 16th ACM international conference on Multimedia, 26-31 October, Vancouver, 159-168

[33] Mikolov, T., Chen, K., Corrado, G., Dean, J. 2013. Efficient estimation of word representations in vector space. In 2013 International Conference on Learning Representations (ICLR), 2-4 May, Scottsdale.

[34] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 5-10 December, Lake Tahoe, 3111-3119.

[35] Yuan, Y., He, L., Peng, L., Huang, Z. 2014. A new study based on word2vec and cluster for document categorization. Journal of Computational Information Systems, 10(2014), 9301-9308.

[36] Salton, G., Wong, A., Yang, C. S. 1975. A vector space model for automatic indexing. Communications of the ACM, 18(1975), 613-620.

[37] Kansheng, S. H. I., Jie, H. E., Liu, H. T., Zhang, N. T., Song, W. T. 2011. Efficient text classification method based on improved term reduction and term weighting. The Journal of China Universities of Posts and Telecommunications, 18(2011), 131-135.

[38] Lan, M., Tan, C. L., Su, J., Lu, Y. 2009. Supervised and traditional term weighting methods for automatic text categorization. IEEE transactions on pattern analysis and machine intelligence, 31(2009), 721-735.

[39] Hall, M. A. 2000. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the Seventeenth International Conference on Machine Learning, 29 June-2 July, Stanford, 359–366.

[40] Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning, 21-23 April, Chemnitz, 137-142.

[41] Kotsiantis, S. B., Zaharakis, I., Pintelas, P. 2007. Supervised machine learning: A review of classification techniques. In Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering, 3-24.

[42] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(2009), 10-18.

[43] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai, 14(1995), 1137-1145.