Pamukkale Univ Muh Bilim Derg, 24(2), 292-301, 2018

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Music emotion classification for Turkish songs using lyrics

Türkçe şarkılar için şarkı sözleri üzerinden müzik duygu sınıflandırması

Ahmet Onur DURAHIM[1], Abide COŞKUN SETIREK[2], Birgül BAŞARIR ÖZEL[3], Hanife KEBAPÇI[4]*

[1,2,3,4] Management Information Systems Department, Boğaziçi University, Istanbul, Turkey.
onur.durahim@boun.edu.tr, abide.coskun@boun.edu.tr, birgulbasarir@yahoo.com, hanifekebapci@hotmail.com

Received/Geliş Tarihi: 29.12.2016, Accepted/Kabul Tarihi: 18.08.2017
doi: 10.5505/pajes.2017.15493
* Corresponding author/Yazışılan Yazar
Research Article/Araştırma Makalesi

Abstract

Music has grown into an important part of people's daily lives. As we move further into the digital age in which a large collection of music is being created daily and becomes easily accessible renders people to spend more time on activities that involve music. Consequently, the form of music retrieval is changed from catalogue based searches to searches made based on emotion tags in order for easy and effective musical information access. In this study, it is aimed to generate a model for automatic recognition of the perceived emotion of songs with the help of their lyrics and machine learning algorithms. For this purpose, first 300 songs are selected and annotated by human taggers with respect to their perceived emotions. Thereafter, Unigram, Bigram and Trigram word features are extracted from song lyrics after performing text preprocessing where stemming of the Turkish words is an essential part. Then, term by document matrices are created where term frequencies and tf-idf scores are considered as representations for the indices. Five different classification algorithms are fed with these matrices in order to find the best combination that achieves the highest accuracy results where recall and precision values are used as comparison metrics. As a result, best accuracy results are obtained by using Multinomial Naïve Bayes classifier where Unigram features are used to create the term by document matrix. In this setting, Unigram features are stemmed by Zemberek Long stemming method, and the index representation is chosen as term frequency. For this combination, obtained recall and precision values are 43.7 and 46.9, respectively.

Keywords: Text mining, Text classification, Sentiment analysis, Music emotion retrieval

Öz

Müzik insanlık tarihinde önemli bir yere sahiptir. Özellikle dijital çağda kişiler tarafından her gün yaratılan ve ulaşılan müzik koleksiyonlarının büyüklüğü ile müziğin önemi daha da artmış ve insanlar müzik içeren aktivitelere daha fazla zaman ayırmaya başlamışlardır. Bununla birlikte, müziğe bilgi geri getirim sürecini kolay ve etkin hale getirmek için yapılan katalog bazlı aramalar duygu tabanlı etiketlere göre aramalara dönüşmüştür. Bu araştırmada amacımız şarkı sözlerine göre bir şarkıdan algılanan duygunun otomatik olarak çıkarıldığı bir model geliştirmektir. Model metin bazlı sınıflandırma için kullanılan makina öğrenmesi algoritmaları ile oluşturulmuştur. Bu amaçla araştırmada 300 şarkı seçilmiş ve bu şarkılar kişiler tarafından hissedilen duygularına göre etiketlenmiştir. Devamında metin ön analizi ile şarkı sözleri Türkçe köklerine ayrıştırılarak Unigram, Bigram ve Trigram kelime özellikleri çıkartılmıştır. Ardından endeksleri terim sıklığı ve tf-idf değerleri olan doküman bazında terim matrisleri yaratılmıştır. Bu matris değerleri 5 farklı sınıflandırma algoritmasına girdi olarak verilerek en yüksek doğruluk sonuçları, hatırlama ve kesinlik metrikleri üzerinden araştırılmıştır. Araştırmanın sonucunda en yüksek kesinlik değeri Zemberek Uzun Kök Ayıştırma Metodu ile Unigram kelime özelliklerine göre ayrıştırılmış ve endeksi terim sıklığına göre belirlenmiş terim bazlı doküman matrisinin Katlıterim Naïve Bayes kümeleyicisinde verdiği görülmüştür. Bu kombinasyonda hatırlama metriği değeri 43.7 iken kesinlik metriği değeri 46.9'dur.

Anahtar kelimeler: Metin madenciliği, Metin sınıflandırması, Duygu analizi, Müzik duygu geri getirim

1 Introduction

Music has grown into an important part of people's daily lives, and as we move further into the digital age in which a large collection of music is being created daily and becomes easily accessible renders people to spend more time on activities that involve music. Everyone may encounter music throughout most routine daily activities such as waking up, eating, working, jogging, swimming, driving, and so forth [1]. As the amount of musical content continues to explode, conventional approaches that manage music pieces based on bibliographic information such as titles, artist names, and genres on a display are no longer sufficient. Hence, music information organization and retrieval has to evolve to meet for the demand for easy and effective information access [1],[2].

Music classification is an essential process for improving music information retrieval (MIR) systems in various media platforms such as Spotify and LastFm, which are the two most widely known music platforms and they have extensive music catalogue. Several approaches such as content-based, context-based, audio-based, etc. are used in order to generate recommendations to listeners [3].

Today, there are several music services which provide large scale music datasets for information extraction and most of the musical content is easily accessible [4]. These music datasets are widely used in order to perform classification of music into predefined categories such as their genres and moods. However, it is essential to assign correct metadata to provide best search results or correct recommendations of multimedia resources [5].

Music listening is a very situational behavior [6]. Lehtiniemi and Ojala [4] stated that the emotional state of the listener is essential for selecting the type of the music for listening. They also argued that specifying the mood of the listener and classifying emotions based on the preferred music by that listener are difficult tasks to accomplish well automatically.

Recently, this demand leads to an increasing interest in the research community to propose and develop tools and algorithms for efficient music organization and retrieval by emotion. It is generally believed that music cannot be composed, performed, or listened to without considering the affection involved. Music information behavior studies have also identified emotion as an important criterion used by people in music seeking and organization [1],[2]. As stated by a

study of social tagging on a popular music website, Last.fm, after the genre and locale tags, mood tag is the third most frequent type of tag assigned to music pieces by online users [7].

As West et al. [8] stated in their study, human beings often use "contextual or cultural labels" for music. Cultural references which lay behind when people define a music piece are changeable. Furthermore, one song can be described mostly with more than a single tag, genre or emotion. Defining them with a single label can be a further limitation.

In one of the music mood detection studies, music mood is classified by asking users to select one mood picture from a set of options rather than a label. This idea is found to be a successful concept in this study and stated to add novel experiences to music listening [4]. According to the study, it is seen as a good way to receive music recommendations from real users based on their mood picture interpretations.

In addition to importance of emotional state in listening music, it is argued that music listening is a very personal behavior [6]. In system development and evaluation, the need for considering the human factors such as preference, activity, and emotion is largely emphasized [9]. Users' involvement and their contribution through non-message-based interactions have become a major force behind successful online communities. Recognition of this new type of user participation is crucial to understanding of the interests of mass population [10]. For these reasons many researchers have called attention to the user-centered design to tag music. In literature, various crowdsourcing tasks are used for human assessment of music mood. Urbano et al. [11] stated that crowdsourcing is a perfectly viable alternative to evaluate music systems without the need for experts.

Users assign rich meanings to music emotion queries, but a music classification algorithm could only retrieve them from the computations' results, which would be shallow in the perspective of users. Consequently, in this study crowdsourcing was used to obtain the emotion tags of various songs. The concept of music emotions from the end-user's perspective was investigated by asking users to choose one emotion cluster from a set of options for various songs. Afterwards, it was tried to formulize a text mining model, which automatically recognizes the emotion of a given music piece from its lyrics. As a side note, the terms "mood" and "emotion" were used interchangeably in this study.

The contributions of this paper are twofold. First, to the best of our knowledge, this is the first study which tries to extract emotions/moods from the Turkish music lyrics. Second, current study considers the use of n-gram features for the purpose of automatically extracting emotions/moods from lyrics, and in the study comparative analysis is conducted with the aim of understanding the effects of applying different stemming methods proposed for Turkish language, term-weighting approaches and classification algorithms on the classification performance.

2 Literature review

2.1 Music Mood Recognition

With the widespread usage of smart phones and personal computers for accessing music and the exploding amount of digital music content available to people necessitate the development of novel algorithms and tools for easy and effective music retrieval. As almost every music piece is created to convey emotion, music organization and retrieval by emotion is a reasonable way of accessing music information [1],[2]. There is a significant amount of study that has been done on the music mood recognition based solely on audio, lyrics, and crowdsourced tags as well as multi modal approaches in where audio, lyrics and tags are used altogether to obtain more accurate and reliable mood classifiers.

Early work on music mood recognition started as a special case of music tagging, by using categorical labels such as happy or sad [12]. Feng and his colleagues [12] used an approach named as Computational Media Aesthetics (CMA), to classify music emotion. In their approach, they assume that composers choreograph the expectation to arise emotion, and performers convert the musical intention into music language to arise emotion. So that, they analyzed music mood on the viewpoint of how music is made. In their scheme, music database is indexed on four labels of music mood, concretely "happiness", "sadness", "anger" and "fear". And three features, relative tempo, the mean and standard deviation of average silence ratio (articulation), are used to classify mood using a back-propagation neural network.

Some web services provide audio decoding of musical features, which are then used as a base for automatic music emotion detection tasks. Echo Nest [13] has offered a web service that provides users a set of musical features, like timbre, pitch, and rhythm. Similarly, the MIR Group of the Vienna University of Technology [14] also made a web service available that returns a set of musical features for a given song such as rhythm patterns, statistical spectrum descriptors and rhythm histograms, and allows the training of self-organizing music maps [15].

In the study by Liu et al. [9], LiveJournal dataset is used to predict user mood. This dataset contains blog articles from the social blogging website LiveJournal. Instead of being collected in a controlled environment, data is contributed by users spontaneously during their regular daily lives. The study offers insights into the role of music in mood regulation and demonstrates how LiveJournal dataset with two-million (LJ2M) articles can contribute to studies on real world music listening behavior [9]. Moreover, a million-scale music-listening dataset was obtained from music related Twitter hashtags in another study [16].

A lyrics based classification technique using n-gram features is proposed by Fell and Sporleder [20]. The novelty of their approach is the varied dimensionality of the lyrics features such as style, song structure and orientation towards the world other than vocabulary and semantics. In order to decide style of the song, a rhyme detection tool is used. Regressive Imagery Dictionary method is applied for semantic evaluation to find the imageries of the lyrics such as conceptual thought and primordial thought.

Hu and Downie [17] present a study comparing music classification techniques using lyrics features and audio features. In this study, in order to find effective features for each specific mood, accuracy of using selected audio and lyrics features a including psycholinguistic lexicon are evaluated among each mood. Most promising accuracy results are achieved using context word (CW) lyrics features where the average accuracy of 61.7% is obtained. Precision or recall values are not provided in the study. In conclusion, lyrics

features are found as the most effective ones in classifying majority of the moods.

2.2 Music classification in non-English languages and text mining of Turkish lyrics

Text mining is a special form of data mining which includes searching in and interpretation of retrieved textual information. Text mining of the song lyrics is a widely used method in MIR and classification. Text mining is a process which generally comprises of text-preprocessing, term-by-document matrix generation and knowledge extraction steps.

Earlier works on lyric analysis for languages other than English are based on lexicon based methods. For instance, Cho and Lee [18] used a manually built lexicon in Korean to extract emotion vectors and recognized moods accordingly. Logan and Salomon [19] categorized stemmed words that are taken from news and lyrics. The aim of their study was to evaluate artist similarities of the songs using lyrics, and they measured similarities based on categorized stems.

Kim and Kwon [21] proposed a method where its strength is claimed to be the feature extraction approach which is adapted to retrieve emotion regarding Korean language's specialties. Such features are measured by emotion condition change, negative word combination, time of emotion and interrogative sentence existence. Howard et al. [22] conduct another study focusing on lyrics in languages other than English for music genre classification problem. In their study, a multilingual setting is considered where songs are written in Spanish and Portuguese. Claiming that traditional text preprocessing techniques may not be suitable for multilingual texts, they run experiments to point out the use of stemming and stop words extraction. As a result, they reported that stopwords removal decrease the accuracy in all classification algorithms.

Türkmenoğlu and Tantuğ [23] performed sentiment analysis of Turkish social media to compare Lexicon and Machine Learning (ML) based methods. In their study, they find out that ML based method performs better than Lexicon based method on both short and long informal texts. In another study, Vural et al. [24] presented a framework for unsupervised sentiment analysis in Turkish text documents. In their work, authors customized SentiStrength sentiment analysis library by translating its lexicon to Turkish and used it for the classification of the polarity of Turkish movie reviews. They achieved 76% accuracy by their proposed technique that is unsupervised and is not specific to the studied problem domain. A more general framework called SentiTurkNet is proposed by Dehkharghani et al. [25] where three polarity scores are assigned to each synset in the Turkish WordNet to indicate its level of positivity, negativity, and objectivity. Using these polarity scores, they achieved 66.7% accuracy for ternary classification of movie reviews.

Zemberek is a Turkish text-preprocessing tool which is the most widely used software library in Turkish text analysis. Although Zemberek provides extensive functionalities for performing different phases of text mining as a whole, such as diacritic restorer for Turkish (deASCIIfier) and part of speech tagger, we extensively focused on word stemming operations. In the literature, several stemming methods have been proposed to find stems of the Turkish words. Tunalı and Bilgin [26] compared Affix Stripping, Fixed Prefix and Zemberek stemming methods and their performances in stemming Turkish texts. As a conclusion, they stated that Zemberek and Fixed Prefix 5 methods are preferable due to their reduction rate [26]. As part of their work, they developed a software program named PRETO which provides several text-preprocessing functionalities among which we have utilized different stemming approaches and assess their effect in music emotion classification.

In text mining, after performing text-preprocessing, term-by-document matrix is generated based on the distribution and occurrences of terms within a set of documents, which are the song lyrics in our case. As for the representation of the indices used in this matrix, the widely used term frequencies and tf-idf values were considered. Tf-idf score is computed as the multiplication of two measures: tf (term frequency) and idf (inverse document frequency). Here, tf represents the frequency of the term within a document (single song lyric), whereas idf indicates how rare is the term among all document set (all song lyrics in the dataset) [27].

2.3 Classification algorithms used for music emotion detection

Mood tag is the third most frequent type of tag assigned to music pieces by online users in Last.fm [7]. In the following, we elaborate on the algorithms used in the literature for genre and emotion classification of music.

Using Gaussian mixture models and diagonal covariance matrices, Tzanetakis and Cook [28] achieved 61% classification accuracy with ten genres. The three features they used for classification were timbre texture, rhythmic content, and pitch content. Hamel and Eck [29] proposed a system that can automatically extract relevant features from audio for a given task. They obtained a classification accuracy of 84.3% on the dataset of Tzanetakis et al. [28] by using deep belief networks and non-linear Support Vector Machine (SVM) classifier. McKay and Fujinaga [30] used feedforward neural networks and k-nearest neighbour classifiers in order to classify the recordings by genre using features based on instrumentation, texture, rhythm, dynamics, pitch statistics, melody and chords. Consequently, for a hierarchical taxonomy consisting of 9 leaf genres, classification accuracies of 98% and 90% were obtained for root genres and for leaf genres, respectively [30].

In Music Emotion Retrieval (MER), emotions are categorized into a number of classes (such as happy, angry, sad, and relaxed), and then selected machine learning techniques are applied to create an emotion classifier [31]. In this respect, several machine learning algorithms have been applied to learn the relationship between music features and emotion labels, such as neural networks [12], support vector machines [32], [33], fuzzy c-means classifier [34], and k-nearest neighbor [35]. Subsequently, models generated through the application of these techniques are used to identify the emotion of a music piece given as the input.

Weninger et al. [36] found that recurrent neural networks outperform both support vector regression (SVR) and feedforward neural networks both in continuous-time and static music mood regression, and achieve an R2 of up to 0.70 and 0.50 with arousal and valence annotations for music mood classification, respectively.

Liu et al.'s [9] study offers insights into the role of music in mood regulation and demonstrates how LJ2M (LiveJournal 2-million) can contribute to studies on real world music listening behavior. They employed the MER models trained from a Last.fm dataset of 31.427 songs, which consider a total number

of 190 music emotion classes. In their study, they have adopted the 12-D EchoNest timbre descriptor as the underlying feature representation, and used support vector machine with the radial basis function kernel. The average accuracy of the 190 binary classifiers is 73.9% in area under curve, according to cross-validation results obtained from the Last.fm dataset.

Measuring similarity of songs or artists using lyrics also attracted attention a lot in the field of text mining. Logan and Salomon [19] used Probabilistic Latent Semantic Analysis method for text analysis of lyrics. Kim and Kwon [21] proposed a lyrics-based emotion classification system based on Partial Syntactic Analysis which reported 58.8% accuracy with their improved emotion features. Bag of Words is another method used for feature extraction of lyrics [22]. However, many researches show that combining text analysis and acoustic analysis provides better results in music classification problem [37].

3 Methodology

In this section, the methodology followed is explained in this research. First, the details on how we gather necessary data for the selected 300 songs are given, and then we elaborate on the emotion tagging process. Thereafter, we try to clarify the text mining analysis process employed in building a classifier to categorize given songs with respect to their perceived/exhibited emotion. The process of our analysis is illustrated in Figure 1, where each step is elucidated in the following subsections.

3.1 Data gathering and preparation

In this phase, 45 Turkish popular music artists were selected from 282 enlisted artists in Turkish Wikipedia page [38]. Thereafter, we have selected 10 to 20 songs from each artist and corresponding lyrics of these songs have been collected with a custom code from the web.

Several problems were encountered with the data collected and had to perform some elimination over this data. Some of the songs were in languages other than Turkish, thus we removed them from the data set. Besides, original versions of songs are intended to be included in the data set. Therefore, if a randomly retrieved song is a remix, acoustic or other kind of adapted versions of the original song, it is also removed from the dataset. Another set of eliminated songs are those which were emotionally confusing songs where it is hard to decide in which mood category they belong to. If people who tagged the song did not agree on a mood, that song is identified as noisy data and excluded from the data set. All songs are tagged for perceived emotions by 3 people. If at least 2 of them agree on the mood category of a song, then it is left in the data set, otherwise it is removed. At the end, we are left with 300 songs in the data set where each one of the four mood categories contains 75 songs. Equal number of songs for each category is selected and tagged with the purpose of avoiding imbalanced learning problem where this approach has been adopted in similar studies [22].

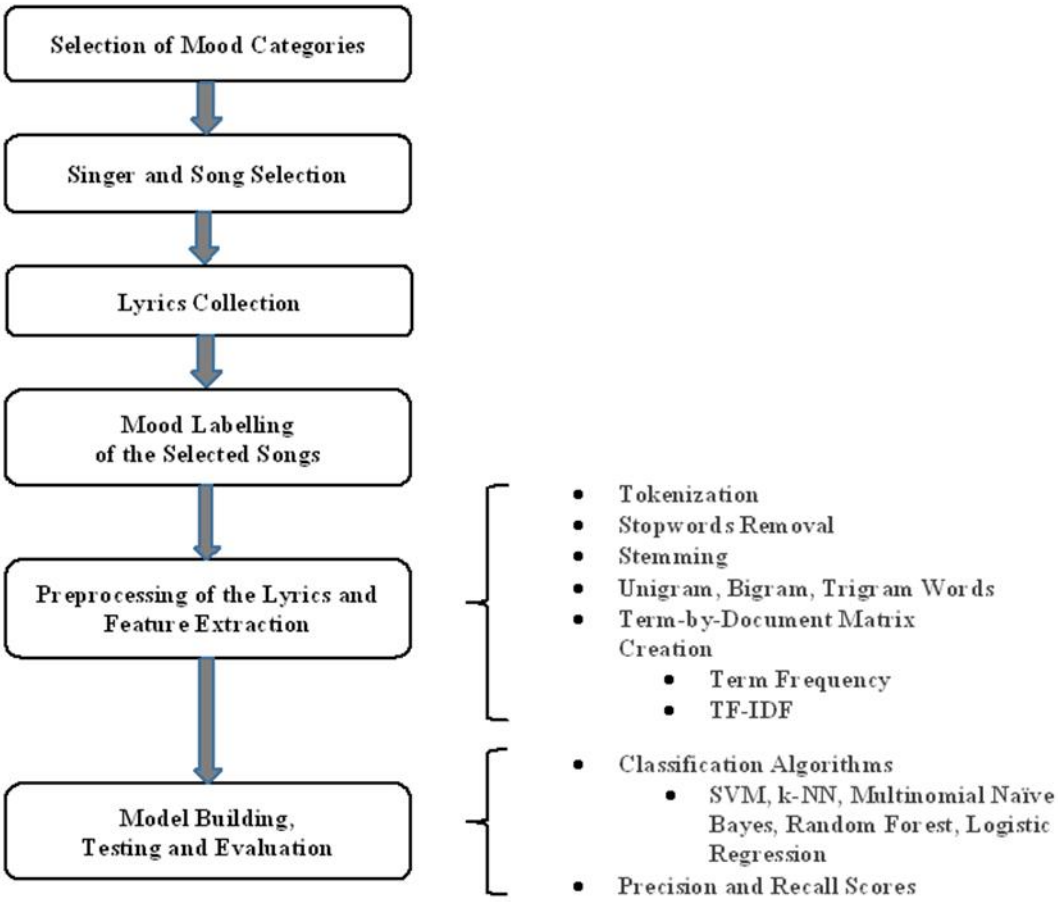

Figure 1: Music mood classification process steps.

3.2 Mood labelling (Emotion tagging)

Russel [39] proposed the circumplex model of affect based on the two dimensional model where the dimensions are "pleasant-unpleasant" and "arousal-sleep". There are 28 affect words in Russel's circumplex models and are shown in Figure 2. Several researchers have adopted a subset of Russel's taxonomy.

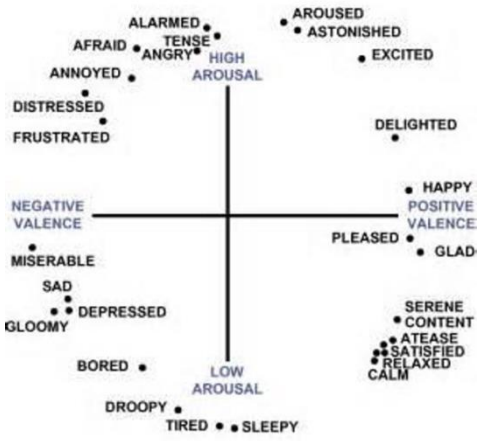

Figure 2: "Circumplex model of affect" (adapted from Russel [39]).

For example, Hu and Downie [17] used all the adjectives including calm, sad, glad, romantic, gleeful, gloomy, angry, mournful, dreamy, cheerful, brooding, aggressive, anxious, confident, hopeful earnest, cynical, exciting. Laurier et al. [40] and Song et al. [41] used happy, sad, angry, and relaxed as mood taxonomy. Patra et al. [34] adapted Russell's [39] model into five clusters with three subclasses.

Based on the literature, we have decided on using 4 emotion categories. There are four clusters in our mood taxonomy with three subclasses, which are shown in Table 1. It is formed by clustering similar affect words of Russels' [39] circumflex model. For example, happy, excited and delighted are placed in the same cluster in order to gather similar songs with respect to their perceived emotions into one group.

Table 1: Four emotion cluster of proposed mood taxonomy.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Happy	*Calm*	*Sad*	*Angry*
Excited	Satisfied	Depressed	Alarmed
Delighted	Relaxed	Gloomy	Tensed

Each one of the selected Turkish songs is tagged by at least 3 different volunteered human taggers, where each one comes from the same socio-economic background and education. Each tagger was a PhD student in department of Management Information Systems. They worked independently and assigned a tag to one of the 4 mood categories. At the end, more than 400 songs are tagged by the human annotators, and among the ones where there is an agreement among at least two of the three annotators, 300 of them are selected in such a way to obtain the class balance between the mood categories as given in Table 2.

3.2.1 Reliability analysis

Based on the annotator agreement results, at least two of the three human annotators can only agree upon the single mood category for 76% of the songs assigned to them just based on their lyrics. Besides, among the songs for which a single mood is agreed upon, only 49% of them are labeled as belonging to the same mood category by all three human annotators. This shows us the difficulty of classifying a given song into a single mood category.

Moreover, the inter-annotator agreement for mood annotation task is examined in order to assess the interrater reliability. Based on the Cohen's kappa [42], we obtained the highest pairwise inter-annotator agreement as 0.61. Besides, inter-annotator agreement based on Fleiss' kappa is moderate at 0.55 level [43],[44].

Table 2: Summary of ground truth data collection.

Emotion	Number of Songs
Happy	75
Calm	75
Sad	75
Angry	75
Total	300

3.3 Preprocessing of the Turkish song lyrics and feature extraction

Lyrics can provide valuable information about the mood of a song. In this respect, to perform classification of the songs into four emotion categories based only on their lyrics, first we extracted lyrics from the database of the music lyrics website "songlyrics.com" [45]. This database provides Java based application programming interface for downloading lyrics with keywords in the form of "track name" + "artist name". For those lyrics of the songs that we could not find in this site, we found them by querying the Google search engine.

In order to accomplish the preprocessing of the Turkish song lyrics, we have utilized PRETO tool, which is designed by Tunalı and Bilgin [26]. This tool was utilized in the Turkish word stemming and stopwords elimination parts. With PRETO tool, one can apply word filtering, such as removal of the words containing less than 3 letters and/or perform de-asciification of words in song lyrics automatically. For the stopwords elimination, we enhanced Turkish stopwords dictionary obtained from python package repository [46]. As a result, 223 stopwords in total were excluded from the lyrics in our analysis.

PRETO tool includes several approaches for stemming of the Turkish words. We have analyzed and compare methods available in PRETO where in their original work, Zemberek method provided the highest quality results in grouping of Turkish words [26].

In our study, all three Unigram, Bigram and Trigram bag-of-words features are extracted from the lyrics after performing stopwords removal, both in their original forms (non-stemmed) and after available stemming procedures applied. Term by document matrices that are created using both original and stemmed words (Affix Stripping, Fixed Prefix, Zemberek and Zemberek Long stemming methods [26] are considered) of the lyrics are fed into supervised learning algorithms to generate corresponding mood detection models. In this respect, both term frequencies and term frequency-inverse document frequency (tf-idf) scores are employed as the index values to compare which one fits best for the Turkish song mood detection task.

In order to generate term by document matrix, first we extracted word stems using PRETO tool. During the extraction phase, we have solved some issues encountered by integrating some Java codes into the source code provided by the authors. Terms, which exist in more than 95% of the lyrics, were eliminated from analysis, since they cannot differentiate the songs from each other.

As mentioned before, we compare two different representations of the indices, term frequencies and tf-idf scores, used in the term by document matrix. All three Unigram, Bigram and Trigram features are considered in this study. First, we generate classifiers just using UNIGRAM word features and then compare the results with the classifiers generated using BIGRAM+ (Unigram and Bigram words together) and TRIGRAM+ (Unigram, Bigram and Trigram words together) word features. Term by document matrices generated by these six possible combinations of representations and n-gram features were fed into different classification algorithms in order to build a classification model for Turkish song moods. At the end, these models are cross-validated based on their accuracies in order to obtain the best mood detection model.

3.4 Classification model building and testing

For the classification model building and testing step, we employed scikit-learn python library [47]. In order to obtain better classification performances, we performed mood detection utilizing different classification methods. Selected algorithms to be considered in this study for the mood detection are; support vector machines (SVM) with linear kernel, of which the libsvm based implementation called SVC method is used, k-Nearest Neighbor method (k-NN) where the best k value is found by the GridSearch method provided in the scikit-learn library, Multinomial Naïve Bayes, Random Forest classifier which contains 100 trees in the forest, and Logistic Regression method.

In order to obtain reliable accuracy performances of these models, and to avoid overfitting, 10-fold cross validation procedure is employed. For the model comparisons, precision and recall values are considered as the accuracy performance metrics.

4 Results

In this section, comparison of the accuracy performances of the models created for music mood detection based on song lyrics are made and explained.

Best accuracy performance is achieved by feeding Multinomial Naïve Bayes classifier with the term by document matrix generated from the unigram stemmed words obtained by applying the Zemberek Long stemming method where the term frequencies are used as the representation for the indices in the matrix. In Table 3, summary results are given for each stemming method considered alongside with the result obtained from using original words. Here, only the results for the combinations that achieve the best accuracy performances sorted by recall scores are given.

As it is seen from the Table 3, Zemberek Long, Zemberek and Fixed Prefix has shown close accuracy performances where use of Zemberek Long stemming method leads to the best precision and recall values.

In Table 4, top 10 accuracy scores obtained from the generated models are given irrespective of the combination of the methods used, again sorted by recall values. It can be inferred from the Table 4 that stemming procedure utilized has significant effect on the performance of the generated classification model where Zemberek and Zemberek Long methods achieved 9 of the top 10 accuracy scores. Besides, only three classification methods, Multinomial Naïve Bayes, SVM and Logistic regression, give rise to top 10 scores whereas accuracy scores obtained from utilizing the other two algorithms, k-NN and Random Forest, are significantly lower as compared to the other methods.

Table 5 shows the accuracy results obtained for each musical mood category by utilizing the selected classifiers where the Zemberek Long stemming method is applied on the lyric words. From Table 5, we can conclude that while for the “Happy” category high accuracy values are obtained consistently. But, best recall scores achieved depends mostly on the classification algorithm used and utilized index representation.

Exceptionally high recall values are obtained for the “Happy” class when k-NN method is used with the term by document matrix created using either BIGRAM+ (Unigram and Bigram words together) or TRIGRAM+ (Unigram, Bigram and Trigram words together) words where the word frequencies are used as the representation for the index values.

In terms of precision, lowest values are always obtained for the “Calm” mood class, whereas best precision scores are obtained for “Happy” and then “Angry” classes, respectively. So, since the lowest values are obtained for the “Calm” mood category, for the best precision and recall scores obtained, one should give emphasis on understanding the reasons for getting low accuracy scores for this category and try to improve the results.

To conclude this section, most frequent (unigram, bigram and trigram) words encountered in the lyrics are given in the following tables. In Table 6, a set of the 10 most frequent unigram words (stems) in the song lyrics is given for each category.

Highly ranked content words seem to have meaningful connections to the categories, such as “aşk/love”, “sev/like”, “kalbi/from heart” in Happy songs. The categories Sad and Angry include similar words which have negative meaning like “ağla/cry”, “kal/stay”, “git/go away” and “yalan/lie”. On the other hand, almost none of the words of “Calm” category carry an emotional meaning except “bırak/leave”.

When we compare the success rates of Unigram, Bigram and Trigram text feature based classification models, we conclude that Unigram words convey more information to the classification algorithms. Table 7 and Table 8 show the 10 most frequent Bigram and Trigram words, respectively.

5 Conclusion

In this study, mood detection of songs with text mining of the song lyrics through bag-of-words approach is investigated. In order to do so, several classification algorithms are examined with various textual features including Unigram, Bigram and Trigram features. Besides, to impose structure on text we generate term by document matrix by utilizing both term frequencies and tf-idf scores and try to understand which representation and feature set fits best for the mood detection problem. Recall and precision scores were used as the performance measures in this study. PRETO tool together with the scikit-learn library were employed for performing preprocessing of the lyrics and creating classifiers, respectively. Three learning algorithms, namely Multinomial Naïve Bayes,

SVM and Logistic regression, best fit for the mood detection task out of the five chosen classification algorithms including k-NN and Random Forest. Besides, Zemberek Long stemming method achieved best accuracy results in terms of both recall and precision values. When we consider n-gram word features, that is to say when the success rates of Unigram, Bigram and Trigram text feature based classification models are compared, we can conclude that Unigram words convey more information to the classification algorithms.

Table 3: Best accuracy scores obtained for each stemming method.

Stemming Method	Index Method	*N*-gram	Classifier	Recall	Precision
Zemberek Long	Term Frequency	Unigram	Multinomial Naïve Bayes	43.7%	46.9%
Fixed Prefix	Tf-Idf Score	Unigram	SVC (linear kernel)	43.7%	42.7%
Zemberek	Tf-Idf Score	Unigram	SVC (linear kernel)	42.3%	44.7%
Affix Stripping	Tf-Idf Score	Bigram	SVC (linear kernel)	37.8%	37.1%
Original (non-stemmed)	Tf-Idf Score	Unigram	Logistic Regression	35.7%	35.4%

Table 4: Combination of methods given for the Top 10 accuracy scores.

Stemming Method	Index Method	*N*-gram	Classifier	Recall	Precision
Zemberek Long	Term Frequency	Unigram	Multinomial Naïve Bayes	43.7%	46.9%
Fixed Prefix	Tf-Idf Score	Unigram	SVC (linear kernel)	43.7%	42.7%
Zemberek	Tf-Idf Score	Unigram	SVC (linear kernel)	42.3%	44.7%
Zemberek Long	Term Frequency	Trigram	Multinomial Naïve Bayes	41.7%	44.3%
Zemberek	Tf-Idf Score	Bigram	Multinomial Naïve Bayes	41.7%	42.8%
Zemberek	Tf-Idf Score	Unigram	Logistic Regression	41.1%	41.1%
Zemberek Long	Tf-Idf Score	Unigram	Logistic Regression	40.5%	40.8%
Zemberek Long	Tf-Idf Score	Unigram	Multinomial Naïve Bayes	40.4%	42.7%
Zemberek	Tf-Idf Score	Trigram	Logistic Regression	40.2%	40.4%
Zemberek Long	Tf-Idf Score	Bigram	Logistic Regression	40.2%	41.0%
Zemberek	Term Frequency	Unigram	Multinomial Naïve Bayes	40.2%	41.5%

Table 5: Accuracy scores obtained for each mood category where Zemberek Long stemming method is utilized.

		ZEMBEREK LONG Stemming Method									
		Recall				Average	Precision				Average
N-Grams	Classification Method	Happy	Calm	Sad	Angry	Recall	Happy	Calm	Sad	Angry	Precision
	kNN	68.0%	28.4%	7.0%	6.6%	**27.5%**	27.2%	27.5%	38.3%	20.0%	**28.3%**
	Logistic Reg.	41.4%	35.9%	40.0%	38.9%	**39.1%**	43.8%	34.3%	38.4%	44.4%	**40.2%**
Unigrams	*SVM*	42.5%	32.1%	29.1%	32.1%	**34.0%**	35.0%	30.6%	29.7%	35.6%	**32.7%**
	MultinomialNB	41.1%	37.1%	55.0%	41.6%	***43.7%***	53.1%	31.5%	46.5%	56.7%	***46.9%***
	RandomForest	47.7%	33.6%	32.5%	23.9%	**34.4%**	36.2%	28.6%	37.8%	42.1%	**36.2%**
	kNN	74.3%	16.1%	0.0%	23.6%	**28.5%**	29.1%	29.1%	0.0%	32.1%	**22.6%**
	Logistic Reg.	47.9%	30.7%	35.2%	40.0%	**38.4%**	46.1%	30.9%	40.1%	40.8%	**39.5%**
Bigrams+	*SVM*	46.6%	36.3%	31.6%	34.3%	**37.2%**	37.6%	33.5%	34.8%	37.7%	**35.9%**
	MultinomialNB	38.9%	32.5%	50.7%	35.7%	**39.5%**	53.2%	25.9%	39.2%	46.4%	**41.2%**
	RandomForest	56.8%	31.1%	33.2%	18.8%	**35.0%**	32.7%	32.9%	52.2%	36.4%	**38.6%**
	kNN	78.0%	22.3%	0.0%	11.8%	**28.0%**	27.5%	27.6%	0.0%	41.7%	**24.2%**
	Logistic Reg.	44.3%	30.2%	31.8%	37.5%	**35.9%**	36.3%	28.4%	47.0%	38.4%	**37.5%**
Trigrams+	*SVM*	47.0%	33.4%	29.3%	35.7%	**36.3%**	36.4%	33.1%	32.9%	40.7%	**35.8%**
	MultinomialNB	40.7%	37.1%	53.6%	35.5%	**41.7%**	52.6%	34.5%	41.0%	49.1%	**44.3%**
	RandomForest	65.5%	24.1%	24.6%	29.6%	**36.0%**	36.6%	26.5%	41.5%	41.9%	**36.6%**
	kNN	45.0%	34.6%	38.4%	19.6%	**34.4%**	41.2%	30.1%	38.9%	29.6%	**35.0%**
	Logistic Reg.	50.9%	31.3%	38.4%	41.4%	**40.5%**	44.6%	35.2%	42.0%	41.3%	**40.8%**
Unigrams	*SVM*	50.2%	30.7%	41.1%	34.8%	**39.2%**	45.9%	26.9%	46.9%	39.6%	**39.8%**
	MultinomialNB	43.2%	35.5%	50.2%	32.7%	**40.4%**	48.7%	31.9%	38.8%	51.5%	**42.7%**
	RandomForest	46.3%	26.4%	38.8%	30.2%	**35.4%**	36.1%	24.2%	43.1%	42.3%	**36.4%**
	kNN	43.0%	31.4%	43.4%	15.7%	**33.4%**	43.1%	23.8%	41.4%	28.7%	**34.2%**
	Logistic Reg.	50.2%	29.3%	41.1%	40.2%	**40.2%**	43.1%	34.9%	43.9%	42.3%	**41.0%**
Bigrams+	*SVM*	45.2%	30.4%	40.5%	38.6%	**38.7%**	41.7%	30.5%	44.9%	39.6%	**39.2%**
	MultinomialNB	46.1%	33.8%	48.2%	29.6%	**39.4%**	48.0%	31.4%	39.8%	40.0%	**39.8%**
	RandomForest	49.6%	26.3%	30.9%	23.0%	**32.5%**	33.6%	27.4%	35.9%	29.9%	**31.7%**
	kNN	46.4%	33.6%	37.7%	18.4%	**34.0%**	37.9%	32.0%	37.7%	28.5%	**34.0%**
	Logistic Reg.	48.2%	29.5%	38.2%	41.6%	**39.4%**	44.7%	32.0%	41.2%	39.9%	**39.5%**
Trigrams+	*SVM*	49.6%	32.0%	36.8%	41.6%	**40.0%**	44.4%	33.8%	38.0%	43.4%	**39.9%**
	MultinomialNB	43.0%	33.6%	46.6%	34.8%	**39.5%**	46.9%	33.1%	39.3%	49.7%	**42.3%**
	RandomForest	53.9%	28.6%	38.8%	25.0%	**36.6%**	34.8%	30.1%	49.2%	34.7%	**37.2%**

Table 6: Top10-ranked unigram word features for each mood category.

Happy	Calm	Sad	Angry
aşk *(love)*	***yok*** *(absent)*	***geç*** *(late)*	***git*** *(go away)*
gel *(come)*	***aşk*** *(love)*	***gel*** *(come)*	***kal*** *(stay)*
sev *(like)*	***geç*** *(late)*	***sev*** *(like)*	***yok*** *(absent)*
geç *(late)*	***sev*** *(like)*	***yan*** *(burn)*	***dünya*** *(world)*
iste *(wish)*	***bil*** *(know)*	***ağla*** *(cry)*	***gel*** *(come)*
bak *(look)*	***gel*** *(come)*	***bil*** *(know)*	***dön*** *(come back)*
yok *(absent)*	***yan*** *(burn)*	***yok*** *(absent)*	***dur*** *(stop)*
git *(go away)*	***gün*** *(day)*	***dön*** *(come back)*	***baş*** *(head)*
kalbi *(from heart)*	***dur*** *(stop)*	***git*** *(go away)*	***ver*** *(give)*
gönül *(heart)*	***bırak*** *(leave)*	***yalan*** *(lie)*	***ağla*** *(cry)*

Table 7: Top10-ranked bigram word features for each mood category.

Happy	Calm	Sad	Angry
tövbe_tövbe *(penitence_penitence)*	***nazar_eyle*** *(evil eye_do)*	***bal_sultan*** *(honey_sultan)*	***tek_baş*** *(alone_head)*
gel_gel *(come_come)*	***eyle_nazar*** *(do_evil eye)*	***ver_ver*** *(give_give)*	***yap_yap*** *(do_do)*
şinanay_şinay *(pean)*	***gel_yan*** *(come_burn)*	***başka_yalan*** *(another_lie)*	***uzak_tutun*** *(far_hold)*
yol_ver *(yield)*	***dem_der*** *(smell_tell)*	***çığlık_çığlık*** *(scream_scream)*	***tutun_uzak****(hold_far)*
nazo_gelin *(special name_bride)*	***aman_kaptan*** *(mercy_captain)*	***damla_gözyaş*** *(drop_tear)*	***oda_birileri*** *(room_somebody)*
halhal_halhal *(anklet_anklet)*	***pazar_eyle*** *(market_do)*	***bekle_bekle*** *(wait_wait)*	***yiğit_yiğit*** *(hero_hero)*
yeni_menajer *(new_manager)*	***iste_yap*** *(wish_do)*	***yarım_kal*** *(half_stay)*	***uzak_uzak*** *(far_far)*
menajer_lazım *(manager_required)*	***geç_kal*** *(late_stay)*	***yan_yan*** *(burn_burn)*	***işçi_yel*** *(worker_breeze)*
kumral_bomba *(brunette_bomb)*	***bırak_git*** *(let_go)*	***kusur_kal*** *(defect_stay)*	***hava_dön*** *(weather_turn)*
yok_yok *(absent_absent)*	***yürü_açık*** *(walk_open)*	***ölüm_başka*** *(die_another)*	***halil_aman****(special-name_mercy)*

Table 8: Top10-ranked trigram word features for each mood category.

Happy	Calm	Sad	Angry
yeni_menajer_lazım *(new_manager_need)*	***nazar_eyle_nazar*** *(evil eye_do_evil eye)*	***ölüm_başka_yalan*** *(death_another_lie)*	***uzak_tutun_uzak*** *(far_hold_far)*
tövbe_tövbe_tövbe *(penitence_penitence_penitence)*	***eyle_nazar_eyle*** *(do_evil eye_do)*	***çığlık_çığlık_çığlık*** *(scream_scream_scream)*	***tutun_uzak_uzak*** *(hold_far_far)*
nazo_gelin_ayak *(special name_bride_foot)*	***yürü_açık_hava*** *(walk_open_air)*	***yürek_yarda_cay*** *(heart_lover_tea)*	***hava_dön_işçi*** *(weather_turn_worker)*
halhal_halhal_halhal *(anklet_anklet_anklet)*	***yan_pazar_eyle*** *(burn_market_do)*	***yarım_keskin_bıçak*** *(lover_sharp_knife)*	***dön_işçi_yel*** *(turn_worker_breeze)*
gelin_ayak_tak *(bride_foot_wear)*	***kaptan_götür_deniz*** *(captain_bring_sea)*	***ver_ver_huzur*** *(give_give_peace)*	***birileri_oda_birileri*** *(somebody_room_somebody)*
ayak_tak_halhal *(foot_wear_anklet)*	***hadi_yürü_açık*** *(let's_walk_open)*	***ver_huzur_ver*** *(give_peace_give)*	***yap_yap_yap*** *(do_do_do)*
şinay_şinanay_hop *(pean)*	***gel_yan_Pazar*** *(come_burn_market)*	***keskin_bıçak_bent*** *(sharp_knife_limb)*	***oda_birileri_oda*** *(room_somebody_room)*
şinanay_şinay_şinanay *(pean)*	***eyle_gel_yan*** *(do_come_burn)*	***kervan_eylem_eylem*** *(camel train_act_act)*	***yalan_yalan_dolan*** *(lie_lie_lies)*
şinanay_hop_şinanay *(pean)*	***aman_kaptan_götür*** *(mercy_captain_bring)*	***kaçak_yarım_keskin*** *(runaway_lover_sharp)*	***tek_baş_tek*** *(alone_head_alone)*
yavru_sına_şinanay *(baby_you_pean)*	***çocuk_büyü_çocuk*** *(child_grow_child)*	***geç_dost_kervan*** *(pass_friend_camel train)*	***baş_tek_baş*** *(head_alone_head)*

6 Limitations and future work

There are some important limitations of this study, some of which are based on the human factors. Music emotion perceptions could not thought apart from the emotional state of people, who are listening to them. Some songs can be classified as Happy, whereas some people sense them as Sad. This situation might cause a limitation of our study and related studies of mood detection. To avoid this problem in this study, songs are annotated by 3 people at the same period of time. Annotators have the same socio economic profile, which although could not represent the whole community, resulting labels can be considered consistent within themselves. Therefore, it is recommended to utilize crowdsourcing for emotion tagging with more taggers to achieve more reliable labelling of the songs.

As a second limitation, we have only benefited from pure n-gram text features. Accuracy performances of the models might be improved by utilizing other syntactic as well as semantic properties of the Turkish language, such as using emotionally representative words, understanding negative expressions for Turkish and utilizing Part-of-Speech tags. In addition, accuracy of the framework can also be improved via incorporating techniques that consider the word orderings.

Moreover, single mood label is chosen for each one of the songs and songs that are not labeled same by at least two of the three annotators are eliminated. Therefore, in order not to lose precious data and improve the accuracy performances, problem can be defined as a multilabel classification problem.

Finally, the different detection models could be applied in different Valence-Arousal Quadrants [48] to get results that are more accurate in estimating a song's emotion. Besides, in a future study, we are planning to integrate text-based features with acoustic features in order to improve classifier performances.

7 Acknowledgment

This research was supported by Bogazici University Research Fund (BAP), Project Number: 15N03SUP2.

8 References

[1] Yang YH, Chen HH. "Machine recognition of music emotion: A review". *ACM Transactions on Intelligent Systems and Technology*, 3(3), 1-30, 2012.

[2] Casey MA, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M. "Content-based music information retrieval: Current directions and future challenges". *Proceedings of the IEEE,* 96(4), 668-696, 2008.

[3] Song Y, Dixon S, Pearce M. "A survey of music recommendation systems and future perspectives". *9th International Symposium on Computer Music Modeling and Retrieval*, London, UK, 19-22 June 2012.

[4] Lehtiniemi A, Ojala J. "Evaluating MoodPic-A concept for collaborative mood music playlist creation". *17th International Conference on Information Visualisation (IV),* London, UK, 15-18 July 2013.

[5] Dulačka P, Bieliková M. "Validation of music metadata via game with a purpose". *8th International Conference on Semantic Systems*, Kristiansand, Norway, 03-06 September 2012.

[6] Okada K, Karlsson BF, Sardinha L, Noleto T. "ContextPlayer: Learning contextual music preferences for situational recommendations". *Asia 2013 Symposium on Mobile Graphics and Interactive Applications*, Hong Kong, China, 19-22 November 2013.

[7] Lamere P. "Social tagging and music information retrieval". *Journal of New Music Research*, 37(2), 101-114, 2008.

[8] West K, Cox S, Lamere P. "Incorporating machine-learning into music similarity estimation". *1st ACM Workshop on Audio and Music Computing Multimedia*, Santa Barbara, CA, USA, 23-27 October 2006.

[9] Liu JY, Liu SY, Yang YH. "LJ2M dataset: Toward better understanding of music listening behavior and user mood". *Multimedia and Expo (ICME)*, Chengdu, China, 14-18 July 2014.

[10] Xia M, Huang Y, Duan W, Whinston A. "Ballot box communication in online communities". *Communications of the ACM*, 52(9), 249-254, 2009.

[11] Urbano J, Morato J, Marrero M, Martin D. "Crowdsourcing preference judgments for evaluation of music similarity tasks". *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*, Geneva, Switzerland, 23 July 2010.

[12] Feng Y, Zhuang Y, Pan Y. "Popular music retrieval by detecting mood". *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 28 July-1 August 2003.

[13] Echonest. "Acoustic Attributes". http://developer.echonest.com/acoustic-attributes.html (01.11.2014).

[14] The MIR Group of the Vienna University of Technology. "Audio Feature Extraction Webservice". http://www.ifs.tuwien.ac.at/mir/webservice (11.01.2015).

[15] Corthaut N, Govaerts S, Verbert K, Duval E. "Connecting the dots: Music metadata generation, schemas and applications". *9th International Society for Music Information Retrieval Conference,* Philadelphia, USA, 14-18 September 2008.

[16] Hauger D, Schedl M, Košir A, Tkalcic M. "The million musical tweets dataset: What can we learn from microblogs". *14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 4-8 November 2013.

[17] Hu X, Downie JS. "When lyrics outperform audio for music mood classification: A feature analysis". *International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 9-13 August 2010.

[18] Cho YH, Lee KJ. "Automatic affect recognition using natural language processing techniques and manually built affect lexicon". *IEICE Transactions on Information and Systems*, 89(12), 2964-2971, 2006.

[19] Logan B, Salomon A. "Music similarity function based on signal analysis". *International Conference on Multimedia and Expo (ICME),* Tokyo, Japan, 22-25 August 2001.

[20] Fell M, Sporleder C. "Lyrics-based analysis and classification of music". *25th International Confereence on Computational Linguistics*, Dublin, Ireland, 23-29 August 2014.

[21] Kim M, Kwon HC. "Lyrics-based emotion classification using feature selection by partial syntactic analysis". *23rd IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, Florida USA, 9-9 November 2011.

[22] Howard S, Silla Jr CN, Johnson CG. "Automatic lyrics-based music genre classification in a multilingual setting". *Thirteenth Brazilian Symposium on Computer Music,* Vitória, Brasil, 31 August-3 September 2011.

[23] Türkmenoglu C, Tantug AC. "Sentiment analysis in Turkish media". *International Conference on Machine Learning (ICML),* Beijing, China, 21-26 June 2014.

[24] Vural AG, Cambazoglu BB, Senkul P, Tokgoz ZO. *A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish*. Editors: Gelenbe E, Lent R. Computer and Information Sciences III, 437-445, London, UK, Springer, 2013.

[25] Dehkharghani R, Saygin Y, Yanikoglu B, Oflazer K. "SentiTurkNet: A Turkish polarity lexicon for sentiment analysis". *Language Resources and Evaluation*, 50(3), 667-685, 2016.

[26] Tunalı V, Bilgin TT. "Türkçe metinlerin kümelenmesinde farklı kök bulma yöntemlerinin etkisinin araştırılması". *Elektrik, Elektronik ve Bilgisayar Mühendisliği Sempozyumu (LECO 2012)*, Bursa, Turkey, 29 Kasım-01 Aralık 2012.

[27] Aizawa A. "An information-theoretic perspective of tf–idf measures". *Information Processing & Management*, 39(1), 45-65, 2003.

[28] Tzanetakis G, Cook P. "Musical genre classification of audio signals". *IEEE Transactions on Speech and Audio Processing*, 10(5), 293-302, 2002.

[29] Hamel P, Eck D. "Learning features from music audio with deep belief networks". *11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, 9-13 August 2010.
[30] McKay C, Fujinaga I. "Automatic genre classification using large high-level musical feature sets". *International Society for Music Information Retrieval (ISMIR)*, Barcelona, Spain, 10-15 October 2004.
[31] Barthet M, Fazekas G, Sandler M. "Music emotion recognition: from content-to context-based models". *International Symposium on Computer Music Modeling and Retrieval*, London, UK, 19-22 June 2012.
[32] Laurier C, Meyers O, Serrà J, Blech M, Herrera P, Serra X. "Indexing music by mood: Design and integration of an automatic content-based annotator". *Multimedia Tools and Applications*, 48(1), 161-184, 2010.
[33] Bischoff K, Firan CS, Paiu R, Nejdl W, Laurier C, Sordo M. "Music mood and theme classification-a hybrid approach". *International Society for Music Information Retrieval (ISMIR),* Kobe, Japan, 26-30 October 2009.
[34] Patra BG, Das D, Bandyopadhyay S. "Unsupervised approach to Hindi music mood classification". *Mining Intelligence and Knowledge Exploration*, Tamil Nadu, India, 18-20 December 2013.
[35] Dewi KC, Harjoko A. "Kid's song classification based on mood parameters using k-nearest neighbor classification method and self organizing map". *International Conference on Distributed Framework and Applications (DFmA),* Jogjakarta, Indonesia, 2-3 August 2010.
[36] Weninger F, Eyben F, Schuller B. "On-line continuous-time music mood regression with deep recurrent neural networks". *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),* Florence, Italy, 4-9 May 2014.
[37] Chi CY, Wu YS, Chu WR, Wu DC, Hsu JJ, Tsai RH. "The power of words: Enhancing music mood estimation with textual input of lyrics". *Affective Computing and Intelligent Interaction and Workshops (ACII 2009),* Amsterdam, Netherlands, 10-12 September 2009.
[38] Vikipedia. "Kategori: Türk Pop Şarkıcıları". http://tr.wikipedia.org/w/index.php?title=Kategori:T%C3%BCrk_pop_%C5%9Fark%C4%B1c%C4%B1lar%C4%B1 (01.11.2014).
[39] Russell JA. "A circumplex model of affect". *Journal of Personality and Social Psychology*, 39(6), 1161-1178, 1980.
[40] Laurier C, Grivolla J, Herrer P. "Multimodal music mood classification using audio and lyrics". *7th International Conference on Machine Learning and Applications*, San Diego, California, USA, 11-13 December 2008.
[41] Song Y, Dixon S, Pearce M. "Evaluation of musical features for emotion classification". *13th International Society for Music Information Retrieval Conference*, Porto, Portugal, 8-12 October 2012.
[42] Cohen J. "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, 20(1), 37-46, 1960.
[43] Landis JR, Koch GG. "The measurement of observer agreement for categorical data". *Biometrics*, 33(1), 159-174, 1977.
[44] Fleiss JL, Nee JC, Landis JR, "Large sample variance of kappa in the case of different sets of raters". *Psychological Bulletin*, 86(5), 974-977, 1979.
[45] Songlyrics Know The Words. "Lyrics". www.songlyrics.com (20.11.2014).
[46] Python Software Foundation. "Python Stopwords". https://pypi.python.org/pypi/stop-words (08.01.2015).
[47] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M. "Scikit-learn: Machine learning in python". *The Journal of Machine Learning Research*, 12, 2825-2830, 2011.
[48] Bigand E, Vieillard S, Madurell F, Marozeau J, Dacquet A. "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts". *Cognition & Emotion*, 19(8), 1113-1139, 2005.