





MOODETECTR: MOOD DETECTION FOR TURKISH LYRICS THROUGH WORD VECTORS

¹Barış ÇİMEN , ²Ahmet Onur DURAHİM 

^{1,2} Bogazici University, Department of Management Information Systems, Istanbul, TURKEY

¹baris.cimen@boun.edu.tr, ²onur.durahim@boun.edu.tr

(Geliş/Received: 15.02.2019; Kabul/Accepted in Revised Form: 02.02.2020)

ABSTRACT: Along with the increasing use of online music platforms, catalogue-based searches have turned into mood-based seeking. In this study, we propose MooDetecTR, a semi-supervised learning framework that employs word vectors for Turkish song mood detection. In this framework, first word vectors are created through a large collection of textual data, which include more than 2.5 million Turkish documents, by using Word2Vec and GloVe algorithms. Subsequently, lyrics vectors are generated through combining already trained word vectors of the words in the lyrics selected for mood detection. Lastly, lyrics vectors are fed into various machine-learning algorithms as features to create models for music mood detection. For comparison, Turkish music mood detection is performed both via traditional bag-of-words model, with TF-IDF weights, and Doc2Vec algorithm. The effects of stemming of the words and stop-words removal on the results are investigated, as well. The best micro-f1 score (54.36%) obtained by the proposed framework is 3.81%, and 2.92% higher (7.54%, and 5.68% relative improvements) than the best score obtained from Doc2Vec and bag-of-words methods, respectively. Consequently, the results obtained show the effectiveness of incorporating word vectors generated using big textual data into Turkish text classification process, which is clearly illustrated by the improved classification performance.

Keywords: Text classification, Feature Generation, Music mood classification, Natural language processing, Word embeddings

MooDetecTR: Kelime Vektörleri Vasıtasıyla Türkçe Şarkı Sözleri için Ruh Hali Tespiti

ÖZ: Çevrimiçi müzik platformlarının kullanımının artmasıyla birlikte, katalog tabanlı aramalar, duygu bazlı aramalara dönüşmüştür. Bu çalışmada, Türkçe şarkıların duygu durum tespiti için kelime vektörlerini kullanan yarı denetimli bir öğrenme çerçevesi olan MooDetecTR önerilmiştir. Bu çerçevede, önce kelime vektörleri Word2Vec ve GloVe algoritmaları ile 2,5 milyondan fazla Türkçe belge içeren geniş bir metinsel veri koleksiyonu kullanılarak oluşturulmuştur. Daha sonra, duygu durum tespiti için seçilen şarkı sözlerindeki kelimelerin, daha önceden eğitilmiş kelime vektörlerinin birleştirilmesiyle şarkı sözleri vektörleri üretilmiştir. Son olarak, oluşturulan bu şarkı sözleri vektörleri, müzik duygu durum tespitinde kullanılmak üzere çeşitli makine öğrenmesi algoritmaları kullanılarak oluşturulan modelleri eğitmek için kullanılmıştır. Türkçe müziklerde duygu durumu tespiti karşılaştırma yapılmak üzere ayrıca, hem TF-IDF ağırlıkları kullanılarak geleneksel kelime çantası modeli ile hem de Doc2Vec algoritması kullanılarak oluşturulan modeller ile gerçekleştirilmiştir. Kelimelerin köklerine ayrıştırılması ve gereksiz kelimelerin kaldırılmasının sonuçlara etkileri de incelenmiştir. Önerilen çerçeve ile elde edilen en iyi mikro-f1 skoru (%54,36), Doc2Vec ve kelime çantası yöntemlerinden elde edilen en iyi skorlardan sırasıyla %3,81 ve %2,92 (%7,54 ve %5,68 nispi iyileştirmeler) daha başarılıdır. Sonuç olarak, elde edilen skorlar, Türkçe metin sınıflandırma uygulamasında büyük metinsel verilerin kullanılması ile oluşturulan kelime vektörlerinin olumlu etkisini artan sınıflandırma başarı performansı ile açıkça göstermektedir.

Anahtar Kelimeler: Metin sınıflandırması, Özellik üretimi, Müzik ruh hali sınıflandırma, Doğal dil işleme, Kelime vektörleri

1. INTRODUCTION

For ages, music has been an indispensable part of our lives. With the development of technology and the rise of Internet, digital music libraries have become easily-accessible. Now, we can enjoy music throughout our daily routine activities such as doing exercise, eating, working, and so on (Yang and Chen, 2012). On the other hand, traditional music management that uses catalogue-based information such as title, artist and genre, has become insufficient while the musical databases have grown at a fast pace. As a result, these expansive databases have revealed the necessity to adapt music information organization and retrieval in such a way to meet the demand for easy and effective information access (Casey *et al.*, 2008; Yang and Chen, 2012).

Music classification is a common task for music information retrieval (MIR) systems of digital music platforms such as Spotify and Last.fm, which have plenty of music catalogues. As an interdisciplinary field of study, MIR aims to expand the understanding and usage of musical data by doing research, and developing applications and tools based on the knowledge acquired from music, computer science, signal processing and cognition (Casey *et al.*, 2008). In order to perform music classification for MIR, the data extracted from songs are processed by using machine-learning techniques to categorize songs automatically. Then, the results obtained through this process may serve as an important tool to help music listeners find the information on any musical content and even find the music they may like but have not been exposed to yet. In this respect, various lyric and audio-based approaches are proposed and implemented in building music recommendation systems to music listeners (Song *et al.*, 2012). These recommendation systems usually rely upon music information retrievals algorithms, especially music classification methods. Music information behavior studies consider emotion as an important criterion in music research and organization (Casey *et al.*, 2008; Yang and Chen, 2012). As indicated by a study of social tagging on Last.fm, emotion or mood tags are the third most common tags after type and local tags annotated to music pieces by online users (Lamere, 2008).

With a wealth of semantic information, lyrics can affect and even alter people's perception of music (Ali and Peynircioğlu, 2006). There has been a substantial amount of research carried out where mood classification of songs is accomplished by considering only the lyrics. Hu *et al.* (2009) applied the bag-of-words (BoW) model in combination with stemming and TF-IDF weighting to generate lyrics features. Then, the features were linked together with a set of spectral features for mood classification with Support Vector Machines (SVM). BoW approach was also utilized in Fell and Sporleder (2014), where n-gram features are used for building mood classification models. Su and Xue (2017), on the other hand, extracted lyric-based features through a method named "bag of sentence".

Thanks to the latest developments in representational learning for natural language processing, new ways for feature learning of discrete objects such as words have begun to gain more importance. Lately, Mikolov *et al.* (2013) have introduced Word2Vec model for learning word representations with greater training efficiency than the previously proposed models.

In a subsequent study, Pennington *et al.* (2014) proposed the GloVe model to obtain word vectors, also named as word embeddings, where it was stated that the Word2vec model suffers from ignoring the global word co-occurrence statistics of a given corpus, and it only investigated the context windows of the words across the entire corpus (Mikolov, Sutskever, *et al.*, 2013; Pennington *et al.*, 2014). GloVe model addresses this problem by considering the global statistics of word co-occurrences in a given corpus in addition to the statistics of local context windows.

The main focus of this study is to utilize word vectors in order to improve the mood classification performance based on lyrics only. Here, our purpose is to incorporate huge amounts of textual data gathered over the internet into a lyrics-based mood classification task. To do so, we first crawled Turkish Wikipedia and news articles together with song lyrics, and used these data to build word vectors. These word vectors are then used to generate document features (vectors for each song lyric) that would be

employed to classify song lyrics to their respective moods. In that respect, we utilized two widely used Word2Vec and GloVe models to create word vectors in a fully unsupervised manner.

Main contributions of this study and our major findings can be summarized as follows;

- Enhancing automatic mood detection of Turkish songs via incorporating word vectors through the collection of huge amounts of textual data.
- Comparison of the performances of Word2Vec, GloVe and Doc2Vec algorithms in music mood classification.
- Analyzing the effects of stemming and stop-words removal for generating word vectors to music mood classification performance.

To our knowledge, this study is the most extensive analysis of the Turkish text classification, in particular for music mood detection, which utilizes more than 2.5 million Turkish documents.

2. RELATED WORK

As the amount of musical content continues to explode, the development of novel algorithms and tools for easy and effective music retrieval becomes inevitable. In this section, related works accomplished in this respect are summarized.

2.1. Music Mood Categorization

Yang and Chen (2012) state that emotion-based music organization and retrieval is a logical way to access music data, for almost every piece of music expresses emotion. One of the earliest studies on the categorization of music moods was performed by Russell (1980), who proposed the circumplex model of affect based on the two-dimensional model where the dimensions were “positive/negative valence” and “high/low arousal”. Song *et al.* (2012) have adopted a subset of Russell’s taxonomy and used happy, sad, angry, and relaxed as mood taxonomy.

Hu and Downie (2010) presented a study that compared selected lyrics features and audio features to find out the most effective features in classifying music moods. The best accuracy results were achieved by using context word (CW) lyrics features where the average accuracy of 61.7% is obtained. In their study, lyrics features were found to be the most effective ones in classifying most of the moods.

2.2. Text Classification in Turkish Language

In the field of Turkish text classification, one of the studies was conducted by Güran *et al.* (2009). Their study basically focused on the n-gram approach, while they performed their experimental studies on documents that were either pre-processed or not. According to the experimental evaluation, using unigram representations and Multinomial Naïve Bayes (MNB) in combination gave the best classification results, which is 95.83%.

In another study, Özgür *et al.* (2004) proposed an anti-spam filtering method developed for Turkish in particular and specific to agglutinative languages, consisting of two separate modules – the Learning Module and the Morphology Module. The study was based on both Artificial Neural Network and Bayesian Network algorithms. They claim that they achieved 90% success in finding spam emails in Turkish on the dataset.

Another study was carried out by Torunoğlu *et al.* (2011) to observe the importance of pre-processing steps in Turkish text classification. In their study, they evaluated a variety of pre-processing methods from stop-words filtering to word weighting on several Turkish datasets. According to the experimental results, the pre-processing step did not create the expected impact on Turkish text classification.

In their study, Uysal and Gunal (2014) suggested that pre-processing has an important role in text classification. Their dataset was based on emails and news written both in English and Turkish, through which they determine how pre-processing methods affect classification of the text documents. They also

researched the ways in which tokenization, stop-word removal, lower-case conversion and stemming processes and their various combinations affect the accuracy of SVM classification algorithm. The study found that some pre-processing methods decrease the accuracy score in classification of text documents, while lower-case conversion and stop-word removal processes improve it.

Gunal (2012) studied the effects of various feature selection approaches on text classification. His studies were based on a hybrid selection method created through the combination of filter and wrapper feature selection methods. Accordingly, features obtained through this method gave more successful results in Turkish text classification, as opposed to the single selection method.

A study conducted by Alparslan *et al.* (2011) aimed to extract information from documents that were classified within the Turkish language. First, word stems were extracted using stemming algorithms which were particularly employed for Turkish text documents. Then, they formed document term matrices through the TF-IDF weighting method with stems obtained after preprocessing. Their method achieved 96.67% accuracy score.

2.3. Word Vectors

In regards to text classification, new approaches have been elaborated in conjunction with the recent developments in deep learning methods for natural language processing. Unsupervised representation learning approaches for creating word vectors, also referred to as word embeddings, have gained much importance and utilized in several text classification tasks. Word vectors constructed via representation learning approaches such as Word2vec (Mikolov, Corrado, *et al.*, 2013; Mikolov, Sutskever, *et al.*, 2013) and GloVe (Pennington *et al.*, 2014) were used in several studies including sentiment analysis (Ouyang *et al.*, 2015), and text mining analysis in Turkish language (Cakir and Guldamlasioglu, 2016).

Word2Vec model suffers from ignoring the global word co-occurrence statistics of a given corpus, and it only considers the context windows of the words across the entire corpus (Pennington *et al.*, 2014). In order to address this problem, Pennington *et al.* (Pennington *et al.*, 2014), proposed GloVe approach, which was also an unsupervised method for learning continuous word vectors similar to word2vec model. GloVe differs from Word2Vec method by taking into account the global statistics of word co-occurrences in a given corpus in addition to the statistics of local context windows.

Some studies compared Word2Vec and GloVe on the performance of the text classification tasks. Zhang and Wallance (2015) conducted a sensitivity analysis of one-layer Convolutional Neural Networks to explore the effect of architecture components on sentence classification. According to their findings, both GloVe and Word2Vec performed relatively same. Similarly, A study by Kenter and Rijke (2015) was proposed to investigate whether determining short text similarity is possible using only semantic features. They generated word embeddings by using of both Word2Vec and GloVe separately to compare results and they found that Word2Vec and GloVe gave relatively same results.

2.4. Paragraph/Document Vectors

Following Mikolov *et al.*'s (2013) approach for learning word embeddings, Le and Mikolov (2014) proposed another algorithm called Paragraph Vector, also known as Doc2vec. This is an unsupervised method for learning distributed representations for documents and sentences. It is very similar to the Word2Vec method. The difference was rooted from the fact that there is a vector for each sentence or paragraph in the Doc2vec method. Le and Mikolov (2014) applied this method for sentiment analysis and achieved 1.3% (or 15% relative) improvement over the best previous result.

3. MUSIC MOOD CLASSIFICATION FRAMEWORK

In this section, we introduce the methodology followed in our study in generating models for music mood classification. In this respect, first, an overview of the steps of acquiring the dataset will be given. And then, mood labeling of selected lyrics, and details of feature extraction and weighting process in generating lyric vectors will be explained. Finally, we concentrate on the model building and evaluation steps where we assess our proposed framework and compare it with traditional BoW approach and Doc2Vec algorithm. Figure 1 summarizes the steps of the mood classification process carried out in this research, where each step is elucidated in the following subsections.

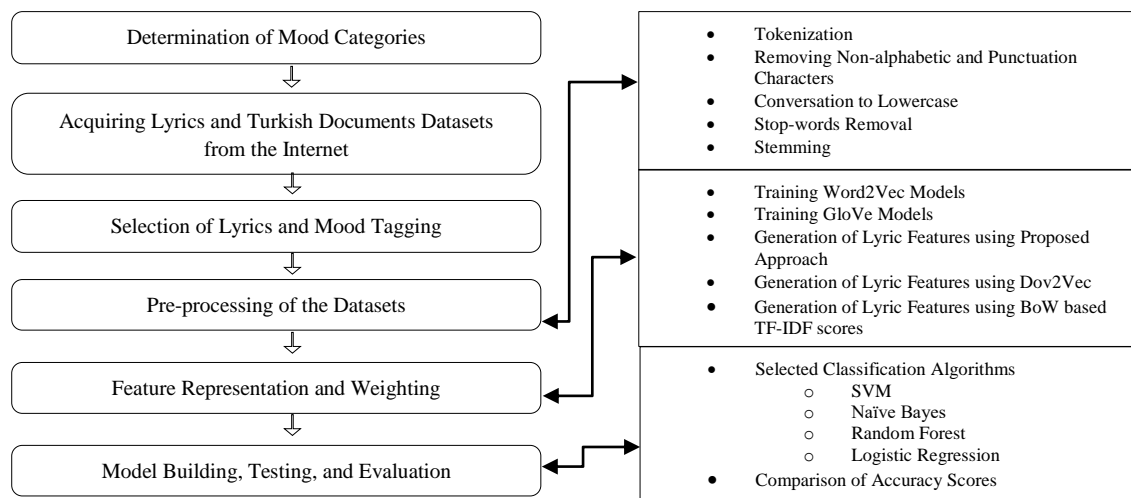


Figure 1. Music mood classification process steps

3.1. Data Collection and Cleaning

Construction of reliable word vectors through neural network models requires huge amounts of textual data. In that respect, we gathered datasets of articles together with the lyrics. Collected lyrics are utilized both in generating the word vectors and in music emotion tagging process. First, more than 300 thousand multi-language lyrics are crawled over the Internet. Then, a language detection tool written in python language called langdetect (Danilák, n.d.) is used to detect the Turkish song lyrics among them, where the number of them reaches over 100 thousand.

After determining and extracting the Turkish song lyrics, we chose a set of popular artists and randomly selected their songs for the manual annotation process and eliminated ones that belong to a remix, acoustic or another adapted version. At the end, we end up with 710 songs in total to be annotated.

Along with the lyrics, a variety of Turkish documents is collected from Turkish Wikipedia and Turkish news websites to be used in the training of word embeddings. Consequently, more than 254 thousand Turkish Wikipedia documents and 2.3 million Turkish news documents are collected. At the end, without any further pre-processing like stemming or stop-words removal, this dataset consists of 570,842,387 tokens, and out of which 2,902,265 of them are unique tokens.

3.2. Mood Tagging

In consideration of the previous literature, in this study, four different mood categories, as happy, calm, sad and angry, are determined and adapted. So, the aim of this study is to automatically classify song lyrics into one of these four mood categories. In this respect, selected song lyrics given to the human annotators, where each one comes from the same socio-economic background, and they are students whose ages ranged from 20 to 35. At least three human annotators label each one of the songs. If the

majority of the annotators agree on a mood category, then corresponding label was set as the mood category for that song. In total, 710 lyrics were annotated and out of which at least two of the three annotators label only 515 of them with the same mood category. Here, 195 of the selected song lyrics were found emotionally confusing where it is hard to decide in which mood category they belong to by the human annotators. These song lyrics were identified as noisy data, and excluded from the dataset. The mood class distribution of the remaining 515 annotated lyrics are given in Table 1.

Table 1. Summary of ground truth class labels

Mood Category	Number of Songs
Happy	76
Calm	78
Sad	253
Angry	108

3.3. Reliability Analysis

Based on the annotator agreement results, we can see that at least two of the three human annotators can only agree upon the single mood category for 73% of the songs assigned to them just based on their lyrics. Besides, among these 515 songs for which a single mood is agreed upon, only 52% of them (which corresponds to 38% over all the annotated songs) are labeled as belonging to the same mood category by all three human annotators. These low percentages in label agreements shows us the difficulty of classifying a given song into a single mood category.

We additionally analyzed the inter-annotator agreement for mood annotation task to evaluate the interrater reliability. Based on the Cohen's kappa (Cohen, 1960), we obtained the highest pairwise inter-annotator agreement as 0.59. In addition to this, inter-annotator agreement computed using Fleiss' kappa is moderate at 0.54 level (Fleiss *et al.*, 1979).

3.4. Data Pre-processing

Pre-processing of the datasets is one of the most important steps in text classification. Tokenization, elimination of stop-words, and stemming are the most commonly-utilized pre-processing methods that are also used in this study.

In this work, we created and used a Turkish stop-word list which consisted of 342 Turkish words. For stemming, we employed a freely-available morphological analyzer for Turkish named TRMorph (Coltekin, 2010). TRmorph is a two-level Turkish morphological analyzer developed for the purpose of high availability and distributed with a license that allows anyone to use and modify it freely for different applications (Coltekin, 2010).

3.5. Learning Word and Lyric Vector Representation

After data preprocessing step, we trained word embeddings with the dataset which consisted of more than 2.5 million Turkish documents. In order to do so, we employed two methods proposed for learning vector space representations of words, namely Word2Vec (Mikolov, Corrado, *et al.*, 2013; Mikolov, Sutskever, *et al.*, 2013) and GloVe (Pennington *et al.*, 2014).

In our study, we considered using Word2Vec's skip-gram model. We trained the Word2Vec model through negative sampling with the parameter value 5 for negative, which specifies how many "noise words" should be drawn. Besides, we set parameter `min_count` to 1 and then 10, in order to analyze its impact on the performance. This parameter is used to ignore all words with total frequency lower than the parameter's value in the entire corpus. In this experiment, size of the sliding window was set to 10,

and word vectors are created with sizes of 100, 200, 300 and 400, which represent the dimensionalities of the feature vectors of the words. Other parameters were kept at their default values.

In the training process of GloVe word embeddings models, we created word vectors of sizes 100, 200, 300 and 400 as well. Rest of the parameters were kept at their default values, whereas the number of epochs was set to 20.

Following the process of training word embeddings, we applied three different approaches in order to generate lyrics vectors for the annotated lyrics. First, we took the average of all words' word vectors that appeared in the song lyrics before filtering stop-words. In the second approach, we carried out the same process after filtering out the stop-words. Finally, we calculated the average of the word vectors by multiplying them by their corresponding TF-IDF scores which were computed over 515 annotated song lyrics with three different threshold values—0.01, 0.001, and 0.00001 respectively. As a result, we obtained feature vectors for song lyrics to be fed into training of the selected machine learning classifiers, which were later used in automatic classification of songs into four mood categories.

For comparison, we also applied Doc2Vec algorithm and traditional BoW approach based on TF-IDF scores to the Turkish music mood detection task. In the BoW approach, to extract n-gram word features, we considered unigram, bigram and trigrams of words.

3.6. Selected Classifiers and Classification

From the viewpoint of machine learning, the objective of text classification is to train classifiers over labelled documents and achieve classification on documents with unknown labels. Considering the problem of high dimensionality, imbalanced class distribution, over-fitting characteristics and previous researches on text classification, this study was focused on four well-known text classification classifiers, namely SVM, NB, Random Forest (RF) consisted of 100 decision trees, and Logistic Regression (LR) and utilized scikit-learn package (Pedregosa *et al.*, 2012).

3.7. Testing and Evaluations

In order to obtain reliable and accurate performances, all the computational results reported in this study used stratified ten-fold cross-validation procedure. Since dataset has imbalanced distribution, hit rate would be a misleading performance criterion. Therefore, we decided on using micro-averaged f1 score as the accuracy performance metric for model comparisons.

3.8. Performance of Proposed Approach

At first, the proposed approach was applied on Turkish music mood detection by using lyric vectors generated by using the word vectors created via Word2Vec and GloVe algorithms separately. Then, the same process was conducted with considering the Doc2Vec algorithm and BoW approach based on TF-IDF scores.

In the first step, the consequential effects of using different word vector dimensions and word counts for word embeddings were investigated. The best result from the experiments is obtained with word vectors with dimension of 100 and setting minimum word count to 10.

In the next step, we perform experiments, and compare performance results for the methods considered in this study, where minimum word count is set to 10 for Word2Vec and GloVe methods. Besides, in further analyses we fixed the lyrics vector size to 100, and therefore make comparisons of the methods examined in this study with lyrics vectors created via running Word2Vec, GloVe and Doc2Vec methods of the same size. The results obtained using different parameter settings for the different methods are given in the Table 2. Although, unigram, bigram, and trigram, were used for the BoW method, only the best-performed n-gram parameter, unigram, is included in the results.

As seen in Table 2, Word2Vec performs better than GloVe for the dataset used in this study. The best score achieved is 52.18%, which was obtained using the RF classifier through the proposed approach

with the lyric vectors computed by averaging the word vectors generated via Word2Vec algorithm. Removal of the stop-words has a minor negative effect on the score obtained from the RF classifier, but it is evident that there is an improvement in the scores of other classifiers. In addition, incorporation of TF-IDF scores into the lyrics vector computation process seems to reduce the score slightly.

The best score obtained from the approaches used in this study is 1.63% higher (or 3.22% relatively) than the best score obtained from the Doc2Vec and 2.44% higher (or 4.90% relatively) than the best score obtained from the BoW approach. This result shows that the proposed method is effective and inclusion of word embeddings in the Turkish text classification process improves the performance.

Table 2. Performance comparisons of the methods considered in the study

Method	Classifier			
	SVM	NB	RF	LR
Proposed Approach using Word2Vec Averages	46.95%	46.71%	52.18%	46.31%
Proposed Approach using Word2Vec Averages + Stop-words Removed	49.12%	49.03%	51.97%	46.54%
Proposed Approach using Word2Vec Averages + TF-IDF (Unigram, Threshold: 0.001)	48.96%	45.10%	52.02%	43.46%
Proposed Approach using GloVe Averages	40.18%	40.65%	51.59%	49.28%
Proposed Approach using GloVe Averages + Stop-words Removed	41.72%	39.20%	50.23%	49.85%
Proposed Approach using GloVe Averages + TF-IDF (Unigram, Threshold: 0.001)	45.41%	38.43%	50.84%	46.21%
Doc2Vec	50.55%	39.02%	50.07%	39.61%
BoW (unigram)	48.93%	47.18%	49.52%	49.74%

In the last step, we investigated the effects of stemming on the classification performance. Due to the fact that Turkish is an agglutinative language, in which word structures are formed by productive affixations of derivational and inflectional suffixes to word roots, stemming is expected to have considerable effect on classification performance. Table 3 shows the comparison of performance results obtained by using stemmed and non-stemmed lyrics in mood classification task:

Table 3 shows that stemming improves the performances of the proposed method and BoW approach. However, stemming does not have a significant effect on the results obtained by Doc2Vec method.

Table 3. Evaluation of stemming on proposed approach using Word2Vec vs. Doc2vec method

		Classifier			
Method	Stemming	SVM	NB	RF	LR
Proposed Approach Using Word2Vec Averages	Stemmed	48.18%	46.37%	54.36%	46.43%
	Original	46.95%	46.71%	52.18%	46.31%
Doc2Vec	Stemmed	48.85%	48.47%	49.88%	47.64%
	Original	50.55%	39.02%	50.07%	39.61%
BoW (unigram)	Stemmed	48.10%	44.96%	51.44%	49.86%
	Original	48.93%	47.18%	49.52%	49.74%

As a result, the proposed method using lyrics vectors calculated via averaging the respective word vectors generated by Word2Vec method achieves 54.36% accuracy score that is 3.81% higher (or 7.54% relatively) than the best score obtained from the Doc2Vec method and 2.92% higher (or 5.68% relatively) than the best score obtained from BoW approach based on TF-IDF scores.

4. CONCLUSION

In this research, a framework that incorporates word embeddings into Turkish text classification process was investigated. To that end, various parameters and settings in generating word embeddings were assessed in Turkish music mood detection task. In order to train the word embeddings, two very popular word embeddings algorithms are employed and their performances are compared in our study; namely Word2Vec (Mikolov, Corrado, *et al.*, 2013; Mikolov, Sutskever, *et al.*, 2013) and GloVe (Pennington *et al.*, 2014). Then, labeled lyrics vectors were created by using these word embeddings. For comparison, Turkish music mood detection is also performed through the Doc2Vec algorithm and popular BoW approach which uses TF-IDF scores. Finally, Turkish lyrics mood detection was conducted by applying the selected machine learning classifier algorithms that use lyrics vectors as features. Micro-averaged F1 scores were used as the performance measures for comparisons due to the imbalanced class distribution experienced in the labeled lyrics dataset. The consequential effects of stemming of words into their roots, and filtering of the stop-words were also investigated. The results of the study show that the score of the proposed approach is 3.81%, and 2.92% higher (7.54%, and 5.68% relative improvement) than the best score obtained from Doc2Vec and BoW methods, respectively.

These results support the fact that word embeddings are effective and efficient representations of the words that may be utilized for text classification purposes, which is consistent with the findings of the previous studies. Besides, training word vectors with a large collection of data, compared to even more powerful approaches like training and utilizing paragraph vectors, one can obtain better results. So, incorporating word embeddings trained with large amounts of textual data into the Turkish text classification process improves its performance.

Music companies and individuals can benefit from the methodology proposed in this study. For example, managing rapidly-growing collections of digital music and building more reliable music recommendation systems have always been a challenge for companies which provide music services, such as Spotify and Apple Music. Using the proposed approach, these companies and individuals may find the kind of music they are looking for and better manage their music collections.

5. LIMITATIONS AND FUTURE WORK

Mood perceptions of songs may vary from one listener to another, depending on their emotional state. Whereas a song can be classified as sad by a listener, another listener may classify the same song as aggressive. This situation may create an obstacle to this research and further researches on related topics. To eliminate this problem in this research, the lyrics were classified by at least three different people separately. On the other hand, it should be noted that reliability of such a research can be increased with the number of annotators. The annotators consisted of young students whose ages ranged from 20 to 35. Although the group of annotators in this research could not represent the whole community, resulting classes can be considered consistent within themselves.

Besides, the success of machine learning applications often increases when the size of the dataset expands. Although annotation is a costly process, the number of annotated lyrics dataset should be kept as large as possible. Utilizing crowdsourcing for emotion classification with more annotators would result in achieving a more reliable classification of the songs. Another way of addressing this problem may be to reformulate the mood classification process as a multi-label classification problem. This way all the labels provided by human annotators might have been included in the classification process. In addition, because a piece of music consists of lyrics and sounds, including lyric features and audio features of songs together in further researches should give better and more reliable results.

To address the problem of imbalance class distribution of the song lyrics and to improve accuracy performances, classifiers may also be generated using preprocessed lyrics dataset where the imbalance is alleviated via utilizing sampling algorithms (He and Garcia, 2009).

Finally, this study is one of the very first studies which incorporate word and document embeddings for the classification of Turkish music. By using of larger labelled and unlabeled datasets, it is possible to obtain more comprehensive and more accurate comparison.

6. FUNDING STATEMENT

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

REFERENCES

- Ali, S. O., & Peynircioğlu, Z. F. (2006). Songs and emotions: Are lyrics and melodies equal partners? *Psychology of Music*, 34(4), 511–534. <https://doi.org/10.1177/0305735606067168>
- Alparslan, E., Karahoca, A., & Bahşi, H. (2011). Classification of confidential documents by using adaptive neuro-fuzzy inference systems. In *Procedia Computer Science* (pp. 1412–1417). <https://doi.org/10.1016/j.procs.2011.01.023>
- Cakir, M. U., & Guldamlasioglu, S. (2016). Text Mining Analysis in Turkish Language Using Big Data Tools. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)* (pp. 614–618). <https://doi.org/10.1109/COMPSAC.2016.203>
- Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges. In *Proceedings of the IEEE* (Vol. 96, pp. 668–696). <https://doi.org/10.1109/JPROC.2008.916370>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Coltekin, C. (2010). A Freely Available Morphological Analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (pp. 19–28).
- Danilák, M. (n.d.). *Langdetect 1.0.7*. Python Package Index.
- Fell, M., & Sporleder, C. (2014). Lyrics-based Analysis and Classification of Music. In *International Conference on Computational Linguistics* (pp. 620–631).
- Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5), 974–977. <https://doi.org/10.1037/0033-2909.86.5.974>
- Günel, S. (2012). Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering and Computer Sciences*, 20(Sup. 2), 1296–1311. <https://doi.org/10.3906/elk-1101-1064>

- Güran, A., Akyokuş, S., Güler, N., & Gürbüz, Z. (2009). Turkish Text Categorization Using N-Gram Words. In Proceedings of the international symposium on innovations in intelligent systems and applications (INISTA) (pp. 369–373).
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hu, X., & Downie, J. S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In Proceedings of the 10th International Society for Music Information Retrieval Conference (pp. 619–624).
- Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *American Music*, 619–624.
- Kenter, T., & Rijke, M. de. (2015). Short Text Similarity with Word Embeddings Categories and Subject Descriptors. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM 2015). <https://doi.org/10.1145/2806416.2806475>
- Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188–1196). <https://doi.org/10.1145/2740908.2742760>
- Lamere, P. (2008). Social Tagging and Music Information Retrieval. *Journal of New Music Research*, 37(2), 101–114. <https://doi.org/10.1080/09298210802479284>
- Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations (ICLR 2013). <https://doi.org/10.1162/153244303322533223>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Retrieved from <https://arxiv.org/pdf/1310.4546.pdf>
- Ouyang, X., Zhou, P., Li, C. H., & Liu, L. (2015). Sentiment Analysis Using Convolutional Neural Network. In 2015 IEEE International Conference on Computer and Information Technology (pp. 2359–2364). <https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.349>
- Özgür, L., Güngör, T., & Gürgen, F. (2004). Adaptive anti-spam filtering for agglutinative languages: A special case for Turkish. *Pattern Recognition Letters*, 25(16), 1819–1831. <https://doi.org/10.1016/j.patrec.2004.07.004>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). <https://doi.org/10.3115/v1/D14-1162>
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation of Musical Features for Emotion Classification. In International Society for Music Information Retrieval Conference (ISMIR) (pp. 523–528).
- Su, F., & Xue, H. (2017). Graph-based multimodal music mood classification in discriminative latent space. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 152–163). https://doi.org/10.1007/978-3-319-51811-4_13
- Torunoğlu, D., Çakirman, E., Ganiz, M. C., Akyokuş, S., & Gürbüz, M. Z. (2011). Analysis of preprocessing methods on classification of Turkish texts. In INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications. <https://doi.org/10.1109/INISTA.2011.5946084>
- Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, 50(1), 104–112. <https://doi.org/10.1016/j.ipm.2013.08.006>
- Yang, Y.-H., & Chen, H. H. (2012). Machine Recognition of Music Emotion. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 40. <https://doi.org/10.1145/2168752.2168754>
- Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. ArXiv Preprint ArXiv:1510.03820. <https://doi.org/10.3115/v1/D14-1181>