Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 1213 – 1227

doi:10.3906/elk-1807-41

Lexicon-based emotion analysis in Turkish

Mansur Alp TOÇOGLU*, **Adil ALPKOÇAK**
Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, İzmir, Turkey

Received: 05.07.2018 • **Accepted/Published Online:** 17.12.2018 • **Final Version:** 22.03.2019

Abstract: In this paper, we proposed a lexicon for emotion analysis in Turkish for six emotional categories happiness, fear, anger, sadness, disgust, and surprise. Besides, we also investigated the effects of a lemmatizer and a stemmer, two term-weighting schemes, four lexicon enrichment methods, and a term selection approach for lexicon construction. To do this, we generated Turkish emotion lexicon based on a dataset, TREMO, containing 25,989 documents. We then preprocessed the documents to obtain dictionary and stem forms of each term using a lemmatizer and a stemmer. Afterwards, we proposed two different weighting schemes where term frequency, term-class frequency and mutual information (MI) values for six emotion categories are taken into consideration. We then enriched the lexicon by using bigram and concept hierarchy methods, and performed term selection for efficiency issues. Then, we compared the performance of lexicon-based approach with machine learning based approach by using our proposed lexicon. The experiments showed that the use of the proposed lexicon efficiently produces comparable results in emotion analysis in Turkish text.

Key words: Turkish emotion lexicon, emotion extraction, lexicon-based emotion analysis, key word-spotting, Turkish emotion analysis

1. Introduction

In today's world, a huge amount of raw text data has been generated with the increasing use of social media applications and communication equipment. As the enormous amount of raw data increases exponentially day by day, new business opportunities arise. The main idea behind these business opportunities is the need to extract meaningful information about a subject such as a topic or a product. This kind of information is very valuable because it provides feedbacks about the related subject whether it is liked or not in the market. In light of these feedbacks, such information is used in many areas such as recommendation systems.

Two similar approaches rise to prominence when it comes to extracting meaningful information about feelings from the raw data. The first is emotion analysis and the second is sentiment analysis also known as opinion mining. While the emotion analysis focuses on different emotion categories such as happiness, sadness, or disgust, the sentiment analysis mainly focuses on positive, negative, and neutral categories. These two analysis types are often referred to as emotion analysis when expressed in Turkish in the literature. However, as already mentioned, one examines a number of emotions expressed in the texts while the other examines the feelings in the texts, i.e. the negative, positive, or neutral situations. In the literature, there are studies both in Turkish and English languages for both emotion and sentiment analyses. However, sentiment analysis is more popular in the literature than emotion analysis for two main reasons. The first one is that since the companies or any foundation require brief information, they demand sentiment analysis from the feedbacks of the people

*Correspondence: mansur.tocoglu@ceng.deu.edu.tr

about the related subject. The second reason is that emotion analysis is a more difficult process because of the diversity of emotion categories.

The studies based on sentiment analysis commonly focus on two main approaches which are lexicon-based and machine-learning approaches [1]. The lexicon-based approach requires preconstructed emotional lexicons to analyze individual terms. Each emotional lexical item given in the lexicon is searched within the text and the weights of each spotted terms are accumulated to a total emotion score [1–4]. On the other hand, the machine-learning approach considers the task of emotion analysis as a text classification problem [5] and it requires a training process which is performed by training the classifier on a labeled text collection [6–8]. Both approaches have their advantages and disadvantages. In the lexicon-based approach, there is no need for labeled training data and decisions taken by the classifier. In addition, the lexicon-based approach is more preferable for dealing with negation and intensification cases [9]. However, this usually requires using a proper emotional lexicon, which is not always available. In addition, it is difficult to take the context into account [10]. In the machine-learning approach, there is no need for a lexicon and in practice it produces higher classification results in terms of accuracy. However, these classification results are only obtained by using a labeled training dataset which is not always available. At the same time, the classifier trained on the texts in one domain cannot perform the same accuracy level in other domains [10–12].

In the literature, many studies are based on developing lexicons for sentiment and emotion analyses in English. The following studies focused on developing a lexicon for the former analysis. Nielsen [13] created a new lexicon that can be used for the sentiment analysis. He made use of Twitter data in the creation of this lexicon. For the values of polarity, he gave positive values between 1 and 5 to the words representing the positive information and gave negative values between –1 and –5 to the words representing the negative information. WordNet [14] is an English electronic dataset containing data for nouns, verbs, adjectives, and adverbs. The data in WordNet is organized as synonym sets which are called synsets. Each synset is composed of synonymous words referring a common semantic meaning. Thelwall et al. [15] developed a library named SentiStrength for lexicon-based sentiment analysis in English. SentiStrength library produces sentiment scores for each word of a given text. To do so, it uses several word lists which are sentimental, booster, idiom, negation, and emoticon lists. SentiWordNet [16] is a lexical resource for sentiment analysis. It provides an annotation based on three numerical sentiment scores (positivity, negativity, and neutrality) for each WordNet synset. Clearly, given that this lexical resource provides a synset-based sentiment representation, different senses of the same term may have different sentiment scores. MPQA [17] is a lexicon containing 8222 terms in total. Each term in the lexicon is labeled with polarity values as positive, negative, and neutral. Additionally, each of them has their intensity values which are strong and weak.

Besides developing lexicons for sentiment analysis, there are also available lexicons for emotion analysis in English as follows: Mohammad and Turney [18] created a new lexicon named EmoLex which contains 14,182 words in total. EmoLex is created by considering eight emotion types which are anger, disgust, fear, expectation, joy, sorrow, surprise, and confidence [19]. Strapparava and Valitutti [20] created a lexicon named WAL in their study. WAL is the abbreviated form of WordNet-affect lexicon. The WAL lexicon was created by using several hundred core words tagged by certain emotion categories. The process at this stage is to find the synonyms of the core words in the WordNet lexicon and assign them the emotional type of the relevant core word. In the resulting lexicon, 1536 words are linked to Ekman's [21] six emotion categories. Stone et al. [22] created a lexicon named GI by tagging 11,788 words with 182 tag categories. There are positive and negative categories as semantic orientation within these 182 tag categories. Apart from this, there are also categories of pleasure,

arousal, feeling, and pain. In another study, Mohammed [23] generated a large dataset called Twitter emotion corpus from Twitter by collecting tweets with hashtags corresponding to Ekman's six emotion categories. Then, he created a new emotion-based lexicon called TEC, containing 11,418 word types, by using the Twitter emotion corpus dataset. In the study [24], an emotion dataset named hashtag emotion corpus was created by using the hashtag structure in Twitter. Here, the process of emotion identification is based on the names of the hashtags. In the first step, this dataset started with six emotions and later became a structure that covers 585 emotions. In the next step, they created a lexicon named hashtag emotion lexicon from the emotion corpus dataset.

The existence of lexicons in Turkish literature is very scarce compared to those in English. Moreover, there are lexicons only for sentiment analysis as follows: Vural et al. [25] created a framework for unsupervised sentiment analysis in Turkish. They created their own lexicon to be used in the framework by translating the lexicon of SentiStrength sentiment analysis library [15]. Akbas [26] focused on opinion mining by extracting aspects of entities on Turkish tweets. The author used a Turkish opinion word list constructed manually and proposed a word selection algorithm to automate finding new words with their sentiment strengths. In another study, the authors [27] developed a Turkish WordNet by translating synsets from WordNet as a part of the Balkanet project. Dehkharghani et al. [28] created SentiTurkNet, which is the first Turkish polarity resource in the literature, by assigning three polarity scores to each synsets found in Turkish WordNet [27]. In another study, Ucan et al. [29] proposed an automated translation approach to construct sentiment lexicons for new languages by using English resources. At the end of their study, they achieved to construct three different lexicons for Turkish. Sevindi [30] translated SentiWordNet [16] lexicon into Turkish and created a Turkish sentiment lexicon with a term size of 12,697.

In the literature, there are very few studies focused on generating sources for emotion analysis in Turkish. To the best of our knowledge, there is no lexicon for emotion analysis in Turkish. However, there are several datasets. Boynukalin [31], generated two datasets for analyzing four emotion categories, joy, sadness, anger, and fear. These datasets are Turkish translations of the ISEAR dataset [32] and collection of 25 children's Turkish fairy tales from several websites. For the classification process, she used three different classification methods which are the naive Bayes, complement naive Bayes, and support vector machine. According to the results, complement Naive bayes gave the best results which obtained the accuracy values of 81.34% for the ISEAR dataset with four classes, 76.83% for the Turkish fairy tales with five classes, and 80.39% for the combination of the two datasets with four classes. The second dataset was generated by Demirci who focused on extracting emotion from Turkish microblog entries [33]. She collected tweets for the six emotions, anger, disgust, fear, joy, sadness, and surprise, using the Twitter search mechanism for hashtags. For each emotion category, Demirci defined hashtags containing the derivatives of each emotion word. As a result, Demirci succeeded in collecting 1000 tweets for each emotion, 6000 tweets in total. To investigate the effects of machine learning classification algorithms, she used the naive Bayes, complement naive Bayes, support vector machine, and K-NN classifiers. According to the results, support vector machine outperformed others by achieving 69.92% classification accuracy. Another dataset was generated by Açıcı [34], where she performed a survey which was carried out among 500 university students from different departments for gathering a dataset for seven emotion categories. The participants were asked to write about a moment of their lives for each emotion categories. As a result of this data compilation process, 3189 documents were collected in total.

In this article, we proposed a Turkish emotion lexicon (TEL)[1] that can be used in emotion analysis

[1]The lexicon can be downloaded from http://demir.cs.deu.edu.tr

in Turkish text for six emotion categories. To the best of our knowledge, it is the first Turkish lexicon which is generated from an original Turkish dataset, TREMO, in the literature [35]. To create the lexicon, we examined the effects of a lemmatizer and a stemmer, two term-weighting schemes, four different lexicon enrichment methods, and a term selection process for lexicon-based emotion analysis, respectively. To evaluate the performance of the lexicon on a different Turkish dataset, we compared the classification results of the lexicon-based approach and machine learning algorithms.

The rest of this article is organized as follows: Section 2 provides detailed information about the materials and methods we used to develop the lexicon. Section 3 presents the results and discussions of experiments implemented to show the performance of the proposed lexicon. The last section concludes the paper providing a summary based on the experimentation we performed and gives a look at furthers studies on this topic.

2. Materials and methods

In this section, we described the materials and methods required to create and examine the lexicon mentioned in this study. We decided to use the TREMO dataset [35] as the material to be used for the generation of the lexicon. Before applying any methods on TREMO, we preprocessed the dataset and found the stem and dictionary forms of each word in the dataset. After the completion of preprocessing, we weighted each term using term frequency, term-class frequency, and mutual information (MI) [36] values. Next, we generated four different lexicons by analyzing each weighted term for biword, concept hierarchy, and the combination of these two approaches. After the creation of the lexicons, we applied term selection phase to decrease the dimension of the corresponding lexicons for efficiency issues.

2.1. Dataset

In this study, we used TREMO dataset, which was created in [35] for emotion extraction in Turkish for six emotion categories which are happiness, fear, anger, sadness, disgust, and surprise. To gather this dataset, a survey was conducted where 27,350 documents were collected from 4709 individuals. A validation process was then performed by 48 volunteered annotators for the elimination of ambiguous documents in terms of emotion categories. After the elimination of 1361 ambiguous documents, the validated TREMO dataset was generated which contains 25,989 documents in total.

2.1.1. Preprocessing

The purpose of the preprocessing is to make the TREMO dataset ready for further operations used in the study. In the first step, we removed punctuation marks, numeric characters, and extra spaces. Next, we performed Zemberek as a stemmer [37] and TurkLemma as a lemmatizer [38] on TREMO dataset and constructed two separate datasets, DS_Z and DS_T, added suffixes of letters Z and T, which are the initials of the approaches used. We then deleted stop-words from the relevant datasets. The statistical data about DS_Z and DS_T are shared in Table 1. The gap of number of terms between both datasets can be explained by the performances of tools Turklemma and Zemberek on foreign nouns, proper nouns, adjectives, and verbs. For example, while Turklemma can achieve extracting the dictionary forms of the terms ”empati”, ”pitbull”, ”aktivite”, ”rutin”, ”türbülans”, ”vertigo”, ”iskender kebap”, ”tiksinç”, and ”tırsmak”, Zemberek cannot provide any stems for these terms. On the other hand, the difference between unique terms of both approaches can be explained by the fact that Turklemma is used as a lemmatizer which focuses on extracting dictionary form of each term and Zemberek tool is used as a stemmer focusing on pruning terms all the way down to roots.

Table 1. Size characteristics of DS_Z and DS_T.

Datasets	# of documents	# of terms	# of unique terms
DS_Z	25,989	120,338	4008
DS_T	25,989	121,539	6289

2.2. Method

After preprocessing the dataset, we focused on creating a lexicon providing promising classification results for lexicon-based emotion analysis in Turkish. To do so, we first constructed the TREMO_LEXBasic which contains all the unique terms within the corresponding dataset. We then used term frequency, term-class frequency, and MI values for weighting each term in the TREMO_LEXBasic. After the creation of the basic lexicon, we used it to generate 3 more lexicons, TREMO_LEXBigram, TREMO_LEXConceptHierarchy, and TREMO_LEXConsolidated, by analyzing each term for bigram, constructing a concept hierarchy manually, and creating the combination of these two approaches for the purpose of enrichment of the lexicon. Afterwards, we applied the term selection method on these lexicons.

2.2.1. Term weighting

Weighting each term in a lexicon plays a crucial role in constructing a proper lexicon. Hence, after preprocessing the TREMO dataset, we focused on weighting each unique term of the corresponding dataset for each emotion category. To do this, we used term frequency, term-class frequency, and MI values. Here, we obtained the MI values by using a feature selection method [36], which measures how much information the presence/absence of a term contributes to make the correct classification decision on a category.

We calculated the weight of each term in the lexicon by using two term-weighting schemes named simple and advanced. The simple scheme calculates the weight of a term by considering only the MI value. On the other hand, the second scheme, advanced, calculates the weight of a term in a more detailed way in order to obtain better classification results. The formulas used in these schemes are shown in Eqs. (1) and (2), respectively.

We calculated the weight of the i^{th} term for c^{th} emotion category, W_i^c, as follows:

$$W_i^c = MI_i^c, \tag{1}$$

$$W_i^c = MI_i^c \times \log_2 tf_i^c \times itcf_i. \tag{2}$$

The values used in these formulas are the MI value, the number of term frequency (tf) taking the logarithm to base two for the corresponding emotion category, and the inverse of term class frequency ($itcf$), which indicates the number of emotion categories containing the i^{th} term.

2.2.2. Lexicon enrichment

In general, term selection is an important step for the sake of both text analysis accuracy and computational efficiency. However, all these trimming processes reduce system performance in terms of recall and precision. Luhn defined the resolving power of words [39], shown in Figure 1. "Accordingly, the high and low frequency terms are not seen as good discriminators and the resolving power or the discrimination capability, is seen to peak at the medium frequency words [40]." In order to add the high and low frequency terms into the selected

term set, we included bigrams for term phrases to decrease frequencies of high frequency terms and construct a concept hierarchy to increase frequencies of low-frequency terms.

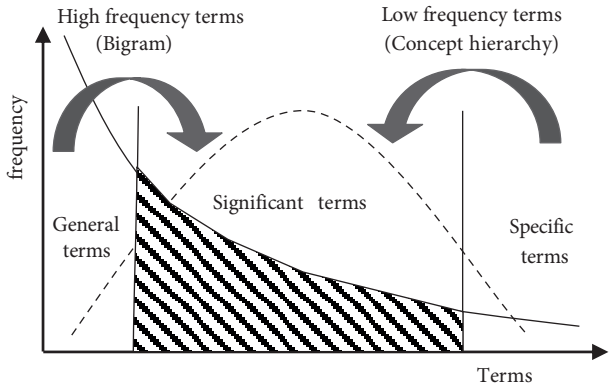

Figure 1. Term frequency diagram [39].

Considering bigram terms simply decreases the high-frequency individual words and increases the chance of selection of these individual words in bigram form. First, we concatenated each word in all documents with the following term and then the term frequency value of each newly concatenated bigram term was calculated. Table 2 shows the overall frequencies of two datasets after including bigrams. The total number of bigram terms in each dataset is high because we added all possible bigram terms to the list irrespective of whether they are meaningful or not. Therefore, we decided to include the first 1000 most repeating bigram terms into the lexicon TREMO_LEXBasic and created a new one named TREMO_LEXBigram. We then calculated the term weights of these newly added bigram terms by using simple and advanced weighting schemes for six emotion categories.

Table 3 presents the first 10 terms with the highest MI values of TREMO_LEXBigram for all emotion categories based on Zemberek (Z) and TurkLemma (TL). There are two main reasons why these terms were given. The first one was to show how the terms changed according to emotion categories. As each emotion category contains the terms related with it, there are also opposite meaningful terms for each category such as happy term which has one of the highest MI values for the angry emotion category. This is because MI value is based on the presence or absence of a term for a category. The second reason was that we tried to show the differences between Turklemma lemmatizer and Zemberek stemmer approaches. For example, "mutlu ol" term is single for happiness emotion category of Zemberek-based lexicon. On the other hand, the same term is represented by three different terms which is the difference of the lemmatization process.

Table 2. Numerical information of the TREMO_LEXBigram types based on the datasets.

Datasets	Total bigram terms	Total unique single terms	Included bigram terms	Total term number
DS_Z	39,858	4008	1000	5008
DS_T	48,154	6289	1000	7289

The main purpose of constructing a concept hierarchy was to push low MI-valued terms into the selected term set. To do this, first, we matched the low MI valued terms consisting of one, two, and three words with their representative terms that will replace them. Table 4 shows some of the terms used for the creation of concept hierarchy for emotion categories fear and sadness. After the replacement process of the low MI valued

Table 3. First 10 terms with the highest MI values of TREMO_LEXBigram for all emotion categories based on Zemberek (Z) and TurkLemma (TL).

Happiness		Fear		Anger		Sadness		Disgust		Surprise	
Z	TL	Z	TL	Z	TL	Z	TL	Z	TL	Z	TL
mutlu	mutlu	kork	kork	öfkelen	öfkelen	üzül	üzül	tiksin	tiksin	şaşır	şaşır
mutlu ol	mutlu oldu	korku	korku	hak	sinirlen	üz	vefat	yemek	koku	bekle	bekleme
ol	mutlu oluru	karanlık	korkut	sinirlen	haksızlık	kork	düşük not	koku	kus	sürpriz	şaşırma
kork	oluru	korku film	karanlık	kork	mutlu	vefat	mutlu	kus	ağız	kork	görün şaşır
kazan	oldu	mutlu	korku film	mutlu	yalan	düşük not	düşük	tükür	mutlu	mutlu	sürpriz
üzül	üzül	gece	mutlu	yalan	üzül	mutlu	öfkelen	kork	böcek	görün	şaşırtı
şaşır	mutlu olu	film	gece	şaşır	tiksin	öl	tiksin	mutlu	tuvalet	tiksin	mutlu
tiksin	öfkelen	şaşır	film	üzül	kork	şaşır	kork	ağız	yer	öfkelen	görün
sevin	kazan	yalnız	üzül	tiksin	şaşır	düşük	şaşır	yer tükür	üzül	üzül	üzül
öfkelen	yüksek not	üzül	yalnız	mutlu ol	al	tiksin	dedem	böcek	tükür	al şaşır	öfkelen

terms with their representatives, we recalculated the simple and advanced weights of each term in the dataset. Thus, we created a new lexicon called TREMO_LEXConceptHierarcy. For example, as one of the representative terms, fear has a new weight 0.246 instead of 0.229 after applying the concept hierarchy enrichment method.

Table 4. Terms used for the creation of concept hierarchy for the emotion categories fear and sadness.

Terms	Concepts	Terms	Concepts
çekinmek (hesitate)	korkmak (fear)	kahretmek (confound)	üzülmek (sadness)
korkunç (terrible)		kahrolmak (be grieved)	
ürkütmek (scare)		burukluk (sourness)	
ürpermek (tremble)		keder (sorrow)	
irkilmek (recoil)		hüzün (sadness)	
ürkmek (boggle)		içerlemek (resent)	
uçurum kenarı (edge of cliff)		canı yakmak (get hurt)	
paniğe kapılmak (panic)		acı olay (tragic event)	
ödü kopmak (be terrified)		keyfi kaçırmak (upset)	
tedirgin olmak (worry)		morali bozulmak (be demoralized)	
kaskatı kesilmek (stiffen)		üzüntü duymak (feel sorry)	
kabus görmek (have a nightmare)		acı vermek (grieve)	
gece vakti (night time)		rahmetli olmak (pass away)	
ödü patlamak (frightened to death)		acı söz (harsh words)	
ölümü çağrıştırmak (conjure death)		vefat etmek (pass away)	

As a result of applying bigram and concept hierarchy methods on TREMO dataset, we managed to obtain two new lexicons. Afterwards, we focused on creating a new lexicon, TREMO_LEXConsolidated, which is the combination of these two lexicons. In this process, we added all the terms to the new lexicon one by one. However, we consolidated the overlapping terms. In other words, we assigned the highest weighting value for the corresponding term. For example, ”mutlu” term has 0.188 MI value in TREMO_LEXBigram and 0.195 MI value in TREMO_LEXConceptHierarcy. In this condition, we assigned the weight of ”mutlu” as 0.195 for TREMO_LEXConsolidated lexicon.

At the end of lexicon enrichment methods, we generated 4 different lexicons for Zemberek and Turklemma individually. Table 5 shows the term numbers of these lexicons based on DS_Z and DS_T datasets.

Table 5. Term numbers of the four lexicons based on DS_Z and DS_T datasets.

Lexicon sets	DS_T	DS_Z
TREMO_LEXBasic	6289	4008
TREMO_LEXBigram	7289	5008
TREMO_LEXConceptHierarcy	6244	3976
TREMO_LEXConsolidated	7235	4966

2.2.3. Key word-spotting technique

In this study, we performed emotion analysis by using a lexicon-based approach, which is an unsupervised learning algorithm. This approach is based on keyword-spotting technique, which uses a previously prepared lexicon to spot terms in a given text.

We assume that the dataset, D, has a set of documents

$D = \{d_1, d_2, ..., d_n\}$

where an arbitrary document d_i is represented with a set of terms,

$d_\mathrm{i} = \{t_1, t_2, ..., t_\mathrm{k}\}$

In the corresponding dataset, the emotion categories, K, has six emotion categories,

$K = \{c_1, c_2, c_3, c_4, c_5, c_6\}$

The key word-spotting technique is a function, which is presented as follows in Eq. (3):

$$E(d_i) = argmax(\sum_{c=1}^{6} \sum\nolimits_{d_{ij} \in L^c} L_j^c w), \tag{3}$$

where d_i indicates the document to be classified using key word-spotting technique, d_{ij} stands for the j^{th} term of the d_i document, L^c indicates the lexicon of the c^{th} emotion category, and w indicates weight value of the j^{th} term for c^{th} emotion category in the lexicon L.

Within the scope of key word-spotting technique, we compared the terms of each document with the terms defined in the corresponding lexicon. If there is a match between the terms, we collected the weights of the matched terms in the lexicon separately for each emotion category and then we calculated the overall scores of each emotion category for each document. Thus, we achieved to determine the new emotion category of the corresponding document by assigning the emotion category with the highest weight value as the new emotion category. Here, if the original emotion category of the relevant document is the same as the newly identified category, no change is made, but if it is different, the emotion category of the related document is replaced with the new one. We redefined the emotion categories of all documents in this way. The ratio of the documents whose original emotion category is not changed in the classification process provides the accuracy value.

2.2.4. Term selection

We applied term selection to choose significant terms for inclusion to lexicon to increase the efficiency in the lexicon-based approach. First, we focused on deciding which lexicon and test dataset to be used. In addition, we also intended to decide which term form to be used, Zemberek or Turklemma. To start with the test dataset, we chose the dataset which was generated by Açıcı [34]. The reason why we made this choice among the others, Boynukalin [31] and Demirci [33], is because the way how the Açıcı's dataset is generated. It is similar to the TREMO dataset which is neither collected from a social media tool nor translated from any language. The categorical distribution of the documents of Açıcı's dataset for this study is shown in Table 6. The test set includes only four emotion categories in common with TREMO, which are happiness, fear, anger, and sadness.

Table 6. Distribution of the test set documents according to emotion categories.

Happiness	Fear	Anger	Sadness	Hate	Shame	Guilty
488	471	471	465	460	417	417

We applied Zemberek as a stemmer and Turklemma as a lemmatizer, and named resulting test sets as TestSet_Z and TestSet_T, respectively. Then, we performed keyword-spotting technique on both sets. In the evaluation experiment, we used four different lexicons with two different weighting schemes to evaluate their

performance in terms of accuracy measures. We then used these results to select an appropriate threshold value to cut lexicons for efficiency issues, based on MI values for each emotion categories individually.

Figure 2 presents the key word-spotting results using simple and advanced weighting schemes for four lexicons on TestSet_Z and TestSet_T, respectively. The results clearly show that using advanced weighting scheme gives higher accuracy values than using simple weighting scheme in all cases no matter which form of test dataset is used. To compare the effects of Zemberek and TurkLemma, we calculated the average accuracy difference between simple and advance weighting scheme, where scores are 0.818% and 2.045%, respectively. In other words, lexicons prepared with TurkLemma generally give higher results. These results can also be explained by the differences between the unique terms of each approach in Table 1, where Turklemma lemmatizer contains the dictionary form of each term and Zemberek stemmer contains only the root forms of each term.

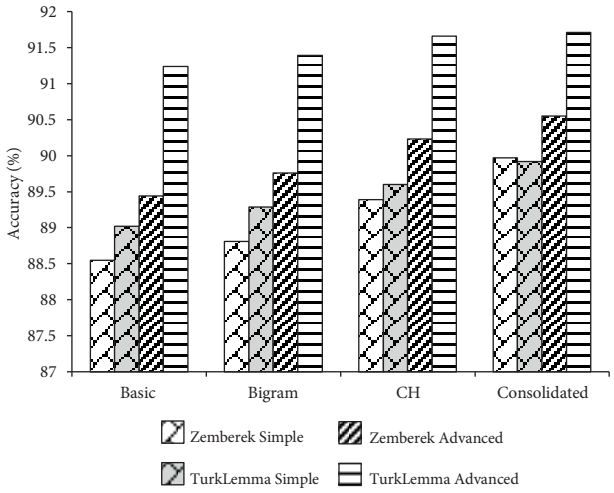

Figure 2. Comparison of Zemberek and Turklemma in terms of accuracy values for different weighing schemes over lexicons. (CH stands for concept hierarchy)

As the four lexicons are compared to each other, The TREMO_LEXConsolidated slightly outperforms the others. This is because it is the combination of both lexicons, TREMO_LEXConceptHierarchy and TREMO_LEXBigram, where both bigram and concept hierarchy enrichment methods are included. On the other hand, there is also a slight difference between both of these two methods, where TREMO_LEXConceptHierarchy outperforms TREMO_LEXBigram. This is related with the generated content of concept hierarchy and bigram term list. According to this two structures, the results might change oppositely.

Based on the results obtained so far, we continued our experiments with TestSet_T and TREMO _LEXConsolidated. At this stage, we weighted the lexicon TREMO_LEXConsolidated by using an advanced weighting scheme. We used this scheme for ranking the most significant terms for each emotion category. We reordered the terms of the corresponding lexicon from the highest to the lowest value and then chose the top n terms as interim lexicons, for each emotion category. We then ran key word spotting using these lexicons and observed the accuracy values. Figure 3 shows the results of this experiment, where the best accuracy value is obtained for $n = 250$. Next, we empirically selected the cut-off value as 0.00091329, which is the weight value of the 250^{th} term of the happiness emotion category. As we applied this value to the other five emotion categories, the term selection value for each category varied as shown in Table 7.

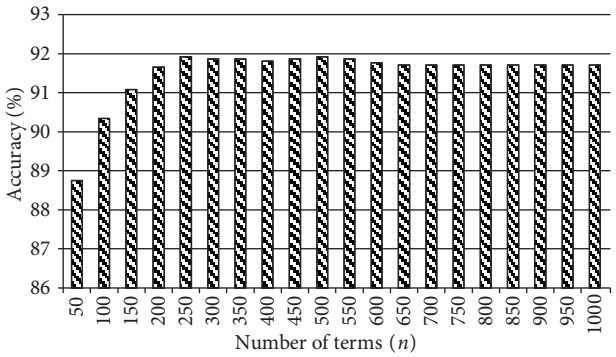

Figure 3. Accuracy values versus number of terms in TREMO_LEXConsolidated.

Table 7. Term counts of TREMO_LEXConsolidated after the term selection process.

Number of terms	Happiness	Fear	Anger	Sadness	Disgust	Surprise
Full terms	7235	7235	7235	7235	7235	7235
Selected term counts	250	214	261	212	228	155

3. Experiments and discussions

After the completion of the term selection process, we implemented two experiments where the Turklemma-based dataset was used. The first experiment focused on the comparison of the performance of selected terms versus full terms where only 3.46% of the overall lexicon was used in keyword-spotting technique. Hence, we ran TREMO _LEXConsolidated on the TestSet_T dataset. We then obtained the overall key word-spotting results using full and selected terms. The results of the experiment in terms of accuracy for full and selected terms are 91.715 and 91.873, respectively. The term selection process slightly increases the overall accuracy value. Although we used only the 3.46% of the overall lexicon in the experiment, we obtained the results without any performance loss. It is clear that reducing the lexicon size positively affects the efficiency of the key word-spotting approach.

Table 8 shows the confusion matrix comparing the emotion categories we obtained at the end of the experiment. In the table, the emotion category includes two columns where left is for full terms and right is for selected terms. Similarly, accuracy columns indicates average accuracy values for each emotion category. Confusion matrix showed that the highest classification result difference is for the happiness emotion category when the selected terms are used instead of full terms. On the other hand, there are no major differences in other emotion categories. For example, the ”anger” category provided slightly higher true positive results for using the selected terms instead of full terms, whereas it is opposite for the categories fear and sadness.

Table 8. Confusion matrix key word-spotting results using TREMO_LEXConsolidated. The results for each emotion category are shown in two columns where the left one is for full terms and the right one for the selected terms.

	Happiness		Fear		Anger		Sadness		Total	Accuracy	
Happiness	474	480	7	4	4	2	3	2	488	97.131	98.361
Fear	9	15	446	444	5	4	11	8	471	94.692	94.268
Anger	17	20	42	39	405	406	7	6	471	85.987	86.2
Sadness	22	27	16	16	14	11	413	411	465	88.817	88.387

In the second experiment, we compared the performance of lexicon-based approach with three machine learning algorithms, support vector machines (SVM), random forest (RF), and naive Bayes (NB), to find out how the lexicon-based approach react against machine learning algorithms.

Tables 9–11 show confusion matrices for the classification results obtained by using three machine learning algorithms, SVM, RF, and NB, respectively. The last columns of these tables provide the individual accuracy values of each emotion category. This experiment showed that SVM has the highest values for three emotion categories, happiness, anger, and sadness. On the other hand, the lexicon-based approach achieved the highest value for the emotion category fear. For SVM, RF, and lexicon-based approach, the happiness emotion category provided the highest accuracy value among the others. The reason why the happiness emotion category became prominent might be because it is easy for people to share their happiness compared to other emotion categories. On the other hand, the anger emotion category obtained the lowest accuracy value for SVM, RF, and lexicon-based approach. This is because anger emotion category is likely to be confused with fear and sadness emotion categories [35].

Table 9. The confusion matrix of support vector machine algorithm.

	Happiness	Fear	Anger	Sadness	Total document	Accuracy
Happiness	479	1	2	6	488	98.156
Fear	10	429	8	24	471	91.082
Anger	11	13	428	19	471	90.87
Sadness	14	5	15	431	465	92.688

Table 10. The confusion matrix of random forest algorithm.

	Happiness	Fear	Anger	Sadness	Total document	Accuracy
Happiness	476	3	0	9	488	97.541
Fear	15	439	3	14	471	93.206
Anger	18	13	417	23	471	88.535
Sadness	17	7	12	429	465	92.258

Table 11. The confusion matrix of naive Bayes algorithm.

	Happiness	Fear	Anger	Sadness	Total document	Accuracy
Happiness	427	5	44	12	488	87.5
Fear	17	414	23	17	471	87.898
Anger	19	15	415	22	471	88.11
Sadness	24	16	53	372	465	80

Table 12 shows the overall evaluation results in terms of accuracy, precision, recall, and f-measure for each classification method. In general, for all evaluation metrics, SVM, RF, and lexicon-based approach are close to each other. However, NB, as a baseline algorithm, falls far behind. For example, the SVM algorithm performs the highest overall accuracy percentage as 93.25% compared to the other three approaches where RF takes the second place with 92.93%, the lexicon-based approach is in third place with 91.71% and NB is in the

fourth position with 85.91%. Besides the overall accuracy value, the other metrics also indicate that SVM is the best classification approach among the others. However, the overall results of the proposed lexicon-based approach is quite close to the SVM results. As this is the case, the proposed lexicon-based approach becomes prominent since there is no training process requirement compared to SVM. In addition, the implementation of a lexicon-based approach is much less complicated to machine learning algorithms used in this experiment.

Table 12. The average results of accuracy, precision, recall, and f-measure of classification algorithms.

	Accuracy	Precision	Recall	F-measure
Lexicon-based	91.71	0.920	0.916	0.917
SVM	93.25	0.933	0.932	0.932
RF	92.93	0.931	0.929	0.929
NB	85.91	0.863	0.859	0.86

4. Conclusion

In this study, we presented a Turkish emotion lexicon (TEL) for emotion analysis in Turkish text for six emotion categories, namely happiness, fear, anger, sadness, disgust, and surprise. We used the TREMO dataset as the source of the newly created lexicon. To generate the terms in the lexicon, we utilized two approaches, Turklemma and Zemberek. The weight of each term was calculated based on term frequency, term-class frequency, and MI values. To improve the term quality of the lexicon, we included bigrams for high-frequency terms and constructed concept hierarchy for low-frequency terms. We then focused on term selection process for efficiency issues. Overall, we investigated the effects of a lemmatizer and a stemmer, two term-weighting schemes, four different lexicon enrichment methods, and a term selection process for lexicon-based emotion analysis.

For evaluation purposes, we performed a set of experiments to find out the best conditions. To do this, we first analyzed the performances of Turklemma lemmatizer and Zemberek stemmer. Out of these results, we came to a conclusion that TurkLemma outperformed Zemberek. This is the difference between using a lemmatizer and a stemmer. This confirms the previous study [41] in literature.

Additionally, we proposed two weighting schemes called simple and advanced. We observed that the advanced scheme produced better results over the simple scheme because of using term and inverse term-class frequency values additively rather than using only MI values for weighting each term in the lexicon. For lexicon construction, we showed that bigrams and concept hierarchy methods provided an improvement in the classification results. Furthermore, we applied term selection for efficiency issues. In the light of the experiments we conducted, we selected the cut-off value as 0.00091329, which is the weight value of the 250^{th} term of the happiness emotion category. We then selected the top terms for each emotion category using the same cut-off value. As a result, we found that the term selection process slightly improved the overall accuracy results for lexicon-based emotion analysis.

In addition to the lexicon-based approach, we also applied three machine learning algorithms, SVM, RF, and NB, on the same test dataset to show how the performance of the lexicon-based approach behaves against machine learning-based approaches. After the experiments we conducted with these algorithms, we observed that our proposed lexicon for emotion analysis produced a comparable result with machine learning algorithms.

All in all, we proposed a lexicon which can be used for emotion analysis in Turkish text. It can be considered effective because we found out comparable results with state-of-the-art performance in the emotion

analysis literature [42]. On the other hand, it is also efficient because we obtained this performance with a very small set of terms, which is less than 4% of the full lexicon set, without performance loss.

For future works, emotion analysis on different datasets, collected from social media tools, by using the lexicon created within this study can be performed. Secondly, automating the construction of the concept hierarchy in lexicon construction can be considered.

References

[1] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Comput Linguist 2011; 37: 267-307.

[2] Ding X, Liu B, Yu PS. A holistic lexicon-based approach to opinion mining. In: Conference on Web Search and Web Data Mining (WSDM); 11-12 February 2008; Palo Alto, CA, USA: ACM. pp. 231-240.

[3] Hu M, Liu B. Mining and summarizing customer reviews. In: International Conference on Knowledge Discovery and Data Mining (KDD2004); 22-25 August 2004; Seattle, WA, USA: ACM. pp. 168-177.

[4] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Association for Computational Linguistics (ACL); 07-12 July 2002; Philadelphia, PA, USA. pp. 417-424.

[5] Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002; 34: 1-47.

[6] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment analysis of twitter data. In: Workshop on Language in Social Media; 23 June 2011; Portland, OR, USA. pp. 30-38.

[7] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Empirical Methods in Natural Language Processing (EMNLP); July 2002; Stroudsburg, PA, USA. pp. 79-86.

[8] Saif H, He Y, Alani H. Alleviating data sparsity for twitter sentiment analysis. In: The 2nd Workshop on Making Sense of Microposts: Big things come in small packages at World Wide Web; 16 April 2012; Lyon, France. pp. 2-9.

[9] Kennedy A, Inkpen D. Sentiment classification of movie and product reviews using contextual valence shifters. Comput Intell 2006; 22:110-125.

[10] Blinov PD, Klekovkina MV, Kotelnikov EV, Pestov OA. Research of lexical approach and machine learning methods for sentiment analysis. Comput Linguist & Intellect Technol 2013; 2: 48-58.

[11] He Y. Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. ACM Trans Asian Lang Inform Proc 2012; 11.

[12] Aue A, Gamon M. Customizing sentiment classifiers to new domains: A case study. In: International Conference on Recent Advances in Natural Language Processing; 21-23 September 2005; Borovets, Bulgaria.

[13] Nielsen FA. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Workshop on Making Sense of Microposts; 30 May 2011; Heraklion, Crete. pp. 93-98.

[14] Fellbaum C. WordNet: An Electronic Lexical Database. London, UK: MIT Press, 1998.

[15] Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment strength detection in short informal text. J Am Soc Inform Sci Technol 2010; 61: 2544-2558.

[16] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: The 7th International Conference on Language Resources and Evaluation; 17-23 May 2010; Valletta, Malta. pp. 2200-2204.

[17] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Lang Resour Eval 2005; 39: 165-210.

[18] Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. Comput Intell 2012; 29: 436-465.

[19] Plutchik R. A general psychoevolutionary theory of emotion. Emotion: Theory, Res, & Exp 1980; 1: 3-33.

[20] Strapparava C, Valitutti A. Wordnet-affect: An affective extension of wordnet. In: The 4th International Conference on Language Resources and Evaluation; 26-28 May 2004; Lisbon, Portugal. pp. 1083-1086.

[21] Ekman P. An argument for basic emotions. Cogn & Emot 1992; 6: 169-200.

[22] Stone PJ, Dunphy DC, Smith MS, Ogilvie DM. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA, USA: MIT Press, 1966.

[23] Mohammad SM. #Emotional tweets. In: First Joint Conference on Lexical and Computational Semantics; 7-8 June 2012; Montreal, Canada. pp. 246-255.

[24] Mohammad SM, Kiritchenko S. Using hashtags to capture fine emotion categories from tweets. Comput Intell 2015; 31: 301-326.

[25] Vural AG, Cambazoglu BB, Senkul P, Tokgoz ZO. A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish. In: 27th International Symposium on Computer and Information Sciences; 3-4 October 2012; Paris, France. pp. 437-445.

[26] Akbas E. Aspect based opinion mining on Turkish tweets. MSc, Bilkent University, Ankara, Turkey, 2012.

[27] Bilgin O, Çetinolu Ö, Oflazer K. Building a wordnet for Turkish. Rom J Inform Sci & Technol 2004; 7: 163-172.

[28] Dehkharghani R, Saygin Y, Yanikoglu B, Oflazer K. Sentiturknet: a Turkish polarity lexicon for sentiment analysis. Lang Resour & Eval 2015; 50: 667-685.

[29] Ucan A, Naderalvojoud B, Sezer EA, Sever H. SentiWordNet for new language: Automatic translation approach. In: 12th International Conference on Signal-Image Technology & Internet-Based Systems; 28 November–1 December 2016; Naples, Italy. IEEE. pp.308-315.

[30] Sevindi BI. Comparison of supervised and dictionary based sentiment analysis approaches on Turkish text. MSc, Gazi University, Ankara, Turkey, 2013.

[31] Boynukalin Z. Emotion analysis of Turkish texts by using machine learning methods. MSc, Middle East Technical University, Ankara, Turkey, 2012.

[32] Scherer KR, Wallbott HG. Evidence for universality and cultural variation of differential emotion response patterning. J Pers & Soc Psychol 1994; 66: 310-328.

[33] Demirci S. Emotion analysis on Turkish tweets. MSc, Middle East Technical University, Ankara, Turkey, 2014.

[34] Açıçı E. Emotion Extraction from Turkish Text. Technical Report, Dokuz Eylul University, İzmir, Turkey, 2012.

[35] Tocoglu MA, Alpkocak A. TREMO: A dataset for emotion analysis in Turkish. J Inform Sci 2018.

[36] Manning CD, Raghavan P, Schütze H. An Introduction to Information Retrieval. Online ed. Cambridge, UK: Cambridge University Press, 2009.

[37] Akın AA, Akın MD. Zemberek, an open source NLP framework for Turkic languages. Struct 2007; 10: 1-5.

[38] Civriz M. Dictionary-based effective and efficient Turkish lemmatizer. MSc, Dokuz Eylul University, İzmir, Turkey, 2011.

[39] Luhn HP. The automatic creation of literature abstracts. IBM J Res & Dev 1958; 2: 159-165.

[40] Ozkarahan E. Database Machines and Database Management. New Jersey, USA: Prentice-Hall, 1986.

[41] Ozturkmenoglu O, Alpkocak A. Comparison of different lemmatization approaches for information retrieval on Turkish text collection. In: 2012 International Symposium on Innovations in Intelligent Systems and Applications; 2-4 July 2012; Trabzon, Turkey. IEEE. pp. 1-5.

[42] Zhang H, Gan W, Jiang B. Machine learning and lexicon based methods for sentiment classification: a survey. In: 11th Web Information System and Application Conference; 12-14 September 2014; Tianjin, China. IEEE. pp.262-265.