Hybrid Emotion Detection with Word Embeddings in a Low Resourced Language: Turkish

Senem Kumova Metin[1], Hatice Ertuğrul Giraz[2]
Department of Software Engineering, Izmir University of Economics, İzmir, Turkey[1]
The Graduate School of Izmir University of Economics, Izmir University of Economics, İzmir, Turkey[2]

***Abstract*—Through natural language processing, subjective information can be obtained from written sources such as suggestions, reviews, and social media publications. Understanding and knowing the user experience or in other words the feelings/emotions of user on any type of product or situation directly affects the decisions to be taken on the regarding product or service. In this study, we focus on a hybrid approach of text-based emotion detection. We combined keyword and lexicon-based approaches by the use of word embeddings. In emotion detection, simply lexicon words/keywords and text units are compared in several different ways and the comparison results are used in emotion identification experiments. As this identification procedure is examined, it is explicit that the performance depends mainly on two actors: the lexicon/keyword list and the representation of text unit. We propose to employ word vectors/embeddings on both actors. Firstly, we propose a hybrid approach that uses word vector similarities in order to determine lexicon words, on contrary to traditional approaches that employs all arbitrary words in given text. By our approach, the overall effort in emotion identification is to be reduced by decreasing the number of arbitrary words that do not carry the emotive content. Moreover, the hybrid approach will decrease the need for crowdsourcing in lexicon word labelling. Secondly, we propose to build the representations of text units by measuring their word vector similarities to given lexicon. We built up two lexicons by our approach and presented three different comparison metrics based on embedding similarities. Emotion identification experiments are performed employing both unsupervised and supervised methods on Turkish text. The experimental results showed that employing the hybrid approach that involves word embeddings is promising on Turkish texts and also due to its flexible and language-independent structure it can be improved and used in studies on different languages.**

***Keywords*—Emotion detection; word embedding; vector similarity; Turkish**

I. Introduction

As a psychological theory, emotion is "a complex psychological condition that includes three separate components: a subjective experience, a physiological response, and a behavioral or meaningful response" [1]. Due to the subjectivity content, people don't all feel same and react to similar situations in the same way. In addition they do not express their feelings alike. Though this variety bring the drawback in emotion analysis, it is not only popular but inevitable to analyze the customer feedbacks and product reviews automatically due to the large amount of data to be processed. This issue is also popular among researchers. It may be stated that there are many studies in the literature for the English language. At the same time, it is promising that the studies for less-resourced languages are increasing every day. We are part of this and in this study, we focused on emotion analysis for Turkish.

The concept of emotion is stated to be the mental state caused by the influence of the environment. In Turkish, though, arousal and intuition are used inadvertently, emotion in English is distinct and is often a means of social knowledge. According to the Oxford dictionary, it is the strong feeling that one feels about his condition, mode and relationship with others. The actions we take, the choices we make and the perceptions we have are affected by the emotions we experience at every moment. In addition, the emotions we feel in same conditions may vary based on several factors such as personality and life experience. Several different theories have emerged to classify and explain the emotions people feel. According to the theory of psychologist Paul Ekman in the 1970s, there exist six basic emotions: happiness, fear, anger, sadness, disgust and surprise. He then expanded the list of basic emotions including some other emotions, such as pride, shame and excitement [2].

Psychologist Robert Plutchik [3] came up with the idea of a "wheel of emotion". According to this theory, several different emotions can be mixed with each other to form an emotion. Just like we mix colors to create other colors. Also, according to Plutchik, more basic emotions act like building blocks. More complex emotions are a mixture of these basic ones.

Though there exists several different approaches in emotion categorization, in our study, we considered six basic emotions proposed by Ekman: happiness, sadness, disgust, fear, surprise and anger. In addition, in this context, the term *emotion lexicon* refers to a list of words and their associations with a set of emotions. Traditionally, in lexicon-based approaches, lexicon words are chosen from arbitrary words of a text resource and are compared to given text units (e.g. sentence, paragraph) in order to assign regarding text unit to one of emotion categories. In comparison operation, a predefined similarity checking procedure is followed of which clearly has a a direct influence on the emotion identification performance.

In this study, our main motivation is to decrease the effort used in emotion detection by a hybrid approach. In this context, we propose the use of well-known word vectors/embeddings presented by Mikolov [4] in two stages of emotion detection process. Word embedding is simply a type of vector-based representation that is obtained by a neural network where a large set of texts is employed. Word embedding method allows words with similar meanings to have similar representations. Thus, it enables to determine semantic relations between words by simple vector based operations such as addition and subtraction. Firstly, we propose to use the pairwise cosine distance between words while choosing lexicon words. In our hybrid proposal, a number of keywords for each emotion is determined similar to keyword-based approaches. But to reduce the effort and bias in keyword-based approaches, this number is set as very limited (in our experiments, for each emotion we determined only two keywords). Following, unlike traditional lexicon-based approaches, all arbitrary words in text are not considered, instead lexicon word candidates belonging to a given emotion category are chosen by similarity measurements to keywords. In other words, words that are assumed to hold a similar semantic content to keywords are chosen as lexicon words. This limited number of lexicon words are labelled by human annotators. Secondly, we propose to employ embeddings again in emotion identification stage. The embeddings of words and text units are compared and similarity scores are used to classify text units to one of emotions.

In our experiments, we built up two new *emotion lexicons* with this hybrid approach and presented three comparison approaches that employ word embeddings in different ways. The paper is structured as follows: Section II is presents a summary for text based emotion recognition and some important works in literature. In Section III, the methodology of the study and in Section IV experimental results are given. Finally, in Section V, we conclude our study.

II. Text-Based Emotion Recognition

Emotions play an important role in human interaction. In today's world, there exist many interaction channels that enable the exchange of emotions and sentiments in different forms such as text and speech. For example, it is common to share our opinions, sentiments and emotions on a product, service or news on different online social platforms. The motivation behind emotion detection studies come from this large amount of online content rich in user opinions, emotions and sentiments. Though a number of people prefer sharing emotions via audio or video files, text is still stated to be the primary choice for people to express their emotions [5]. This made research on emotion extraction from text a popular topic in computational linguistics. As a result, in literature there are many surveys (e.g. [5] [6] [7] [8] [9] [10]) that discuss computational approaches in emotion recognition from different points of view.

Emotion detection methods are generalized into four categories in [5] but there also exists surveys where the first two categories are merged in one. In the categorization of [5], the first approach is stated as keyword-based emotion detection. In keyword-based studies, the main goal is to find out patterns similar to a list of predetermined emotion keywords. Emotion keywords are chosen based on a specific emotion model and the list of emotion words can be improved by the use of online tools and different data resources. For example, in [11], WordNet Affect that is an extension to WordNet [12] is employed. The first weakness of keyword-based emotion detection is the word matching (keyword spotting) that is stated to be simply finding occurences of keywords in the given text [6]. The word matching ignores the semantic relations among words. For example, if a synonym of a keyword is used in text, it is ignored erroneously. The second weakness is the bias while the keywords are determined. For example, a keyword that represents a specific emotion ideally may be a rarely used one in language. The second category is lexicon-based emotion detection that is named as lexical affinity in [6]. This category is stated to be strongly related to, even if an extension to, keyword-based approach [6]. In lexicon-based approach, there exists a knowledge-base with text labeled according to emotions. Though the methods to classify the text to one of the emotions is same with keyword-based approaches, in this category an *emotion lexicon* is utilized instead of a keyword list. The words in lexicon are not directly related to emotions. In other words, the lexicon words are chosen from arbitrary words in given input texts. The words and/or sentences in this predetermined set of texts is labelled commonly by crowd sourcing or multiple annotators and a weight value for each emotion is provided for these arbitrary words. EmoSenticNet (ESN) [13] [14] [15] [16], National Research Council Canada (NRC) Emotion Lexicon [17] [18] and DepecheMood (DPM) [19] are some well-known lexicons where different weighting approaches are utilized. For example, in DepecheMood, tf-idf weighting method is applied to obtain weight values for a set of 8 emotions (afraid, amused, angry, annoyed, happy, inspired, sad, dont care) for each arbitrary word in input texts. The disadvantages of lexicon-based approaches are two-folds. The first is that since the lexicon words are arbitrary words in input texts, the lexicons do not perform well if these words are not occuring in testing texts or if the word holds some meaning other than given in input text. The second disadvantage is that the weight values are biased toward corpus specific genre of texts [6]. The third category in emotion detection covers machine learning methods. Both supervised and unsupervised methods are included in this category where a classifier is trained with a part of dataset and is then used to test the rest of the set. In supervised approaches the dataset or at least a part of the dataset is to be labelled. The studies [20] employing LSTM-based deep learning, [21] using support vector machines, [22] and [23] running unsupervised learning methods, are a few of current works where various machine learning methods are used. The last category is given as hybrid approches of emotion detection where any two or all three approaches are combined to improve the performance or to cope with the disadvantages and weaknesses of previous approaches.

In the literature, though the studies on Turkish is limited compared to other more resourced languages such as English, there are a number of works where new data resources and/or detection methods are presented for sentiment and emotion analysis. For example, Dehkharghani et al. created *SentiTurkNet*, which is one of the pioneering Turkish polarity data set [24]. In *SentiTurkNet*, three polarity scores are assigned to each synset in the Turkish WordNet [25], indicating its positivity, negativity, and objectivity (neutrality) levels in order to be used mainly in sentiment analysis studies. In another study, [26] constructed a system for extracting aspect-based sentiment summaries on Turkish tweets. In [26] a Turkish opinion word list is constructed manually and a word selection algorithm to automate finding new words with their sentiment strengths is proposed. In [27], utilizing a set of 2000 movie comments, emotion-thought analysis is conducted using classification algorithms (e.g. Naive Bayes, center based classifier, multilayer detection and support vector machines) in order to distinguish positive and negative emotions. In [28], an automatic translation approach is presented that creates a sentiment lexicon for a new language from available English resources. In this approach, an automatic mapping is generated from a sense-level resource to word-level by applying a triple unification process. This process produces a single polarity score for each term by incorporating all sense polarities. The major idea is to deal with the sense ambiguity during the lexicon transfer and provide a general sentiment lexicon for languages like Turkish, which do not have a freely available machine-readable lexicon. In [29], a hybrid system is proposed for Turkish sentiment analysis, which combines the lexicon-based and machine learning (ML)-based approaches.

The first Turkish dataset that includes labels for multiple emotions was presented in [30]. In this work, 6000 tweets in total were collected for six emotions (joy, fear, anger, sadness, disgust and surprise) using the Twitter search mechanism for hashtags. The dataset was manually labelled and was utilized in classification experiments. It is reported that support vector machine performing better than the other supervised algorithms achieved a classification accuracy of 69.92%. In [31], ISEAR dataset [32] was translated to Turkish and merged with a set of Turkish fairy tales generating two datasets to be used in analyzing four to five emotion categories. It is reported that ISEAR dataset classification with four classes reached 81.34%, fairy tales dataset classification with five classes reached 76.83% accuracy values by using complement Naive Bayes classifier.

In [33], the Turkish lexicon TREMO (Turkish Emotion Lexicon) dataset that is also used in our study was employed to measure the performance of two different weighting schemes in which term frequency, term class frequency and mutual knowledge values were taken into account. Further information on the lexicon TREMO [34] covering six emotional categories (happiness, fear, anger, sadness, disgust and surprise) is given in next sections. In [33], this lexicon was also enriched using the bigram and concept hierarchy methods and the performance of the lexicon-based approach was compared with supervised machine learning-based approaches. In [33], the experiments are performed on a limited number of testing texts and the performance is measured for four main emotions. Mainly, performance change among emotions is discussed and the overall performance is reported to be in range [85.91% 93.25%]. In [35], deep learning methods are employed to classify Turkish tweets to six emotions. A Turkish tweets dataset is built and annotated automatically using a lexicon-based approach. In [35], convolutional networks is observed to generate the highest accuracy score of 74%.

III. Methodology

In this study, we consider emotion detection as a staged process. The first stage involves the construction of *emotion lexicon*. The second is building text unit representation that will be named as *emotion vector* of sentence. The last stage is labeling the text unit to either one of six emotions by supervised and unsupervised approaches.

In our experiments, the text unit is set to sentence and we run both unsupervised and supervised methods assuming that each sentence must be assigned to one of six emotion categories (happy, anger, rear, sadness, disgust, surprise). Two lexicons named as *Lexicon1* and *Lexicon2* are built by proposed word vector similarity measure. The word vectors are CBOW (continuous bag of words) word embeddings [4] obtained from Wikipedia.

The procedure to build lexicons employing word embeddings and regarding statistical information on proposed lexicons *Lexicon1* and *Lexicon2* are given below. In addition, we also introduce the base set *Lexicon3* that is used to compare the performance of our proposed lexicons.

Lexicon1: To build up *Lexicon1*, firstly keywords are determined. To choose keywords, Turkish names of emotions (*mutluluk*-happiness, *öfke*-anger, *korku*-fear, *iğrenme*-disgust, *şaşkınlık*-surprise, *üzüntü*-sadness) are considered. The words that begin with roots of these names are retrieved and their occurence frequencies are calculated in Wikipedia dataset. The words that have the highest two frequencies are assigned as the keywords of regarding emotion. Following, for each key, ten closest words are determined by measuring the cosine similarity of their word embeddings to the keyword's embedding. Finally, the keyword and the set of ten closest words are packed to form the word list. In Table I, the couples of word lists constructed for each emotion category are given in columns. The second row in Table I includes the keywords of the word lists given in columns. Later, word lists are given to the survey participants. Each participant examine two alternative word lists (list lenght = 11 words) for each emotion and decided the list that represents the regarding emotion better. In Table I, the columns with bold values refer to the word lists that are chosen to be included in *Lexicon1*. In *Lexicon1*, a total of 66 lexicon words, 11 words for each emotion, is obtained.

Lexicon2: *Lexicon2* is built employing the chosen keywords in *Lexicon1*. In *Lexicon2* in order to improve the word lists, twenty closest words to each keyword are determined by measuring cosine similarity of embeddings. Following, a survey is conducted with five undergraduate students to choose ten emotion words from the given set of twenty words for each emotion, the participants examined 126 words in total. Based on the majority of votes, a set of ten words is defined for each emotion. In Table II, *Lexicon2* consisting of 66 words in total is given, it is examined that 29% of the words in list is different than *Lexicon1*.

Lexicon3: *Lexicon3* is the *emotion lexicon* that will be used as the base set in our experiments. This set was presented in [33] and it was built by a survey of 5000 participants where the participants were asked to share their memories and experiences as text for 6 emotional categories. 27.350 documents were collected from the participants and emotion words were chosen from these documents initially. Later, these documents and associated emotion labels were compiled to form TREMO dataset [34]. The lexicon words were recompiled in order to improve the representativeness of emotions. TREMO set is then used to obtain weight for each lexicon word. The weights are stated to be mutual information (MI) values of emotion words in [34]. As a result of this process, *Lexicon3* was built involving 1320 words together with their weights.

TABLE I. WORD LISTS USED TO BUILD *Lexicon1*

HAPPINESS		**FEAR**		**ANGER**		**SADNESS**		**DISGUST**		**SURPRISE**	
mutluluk	mutlu	**korku**	kork	**öfke**	öfkelenir	**üzül**	üzülür	**iğrenç**	iğrenme	**şaşır**	şaşırır
sevgi	mutsuz	**dehşet**	geldükçe	**kızgınlık**	dinlemez	**üzülür**	üzül	**ahlaksız**	oburluk	**sanır**	sanır
iyilik	sevgil	**umutsuzluk**	umr	**umutsuzluk**	üzülür	**affeder**	affeder	**şehvet**	samimiyetsizlik	**üzgü**	zanneder
duygus	üzgü	**korkus**	üzülecek	**şaşkınlık**	affeder	**pişma**	öfkelenir	**gülünç**	beslemez	**şaşırır**	sinirlenir
ıstırap	hasret	**öfke**	dinley	**öfkes**	sinirlenir	**inandırır**	sinirlenir	**aptal**	haset	**korkmuş**	inanmaz
özlem	mutluluk	**çaresizlik**	gülümser	**acı**	angelica	**sevinir**	sevinir	**korkak**	ilişkiselkarşılıklı	**sinirli**	anlar
duygu	sevmek	**vahşet**	pesimis	**üzüntü**	iago	**sinirlenir**	angelica	**acayip**	algılamış	**uyandırır**	şüphelenir
ölümsüzlük	sevdik	**korkunç**	duyamaz	**kıskançlık**	setsuna	**öfkelenir**	setsuna	**delilik**	nefsaniyet	**sinirlenir**	üzülür
tutku	hayaller	**zombi**	polyphemos	**çaresizlik**	öldüresi	**inanmaz**	şaşırır	**alaycı**	seslenmek	**hissettik**	söyley
pişmanlık	aşkı	**duygusallık**	ağlamak	**pişmanlık**	şaşırır	**üzüle**	zanneder	**korkar**	kabullenir	**aptal**	üzgü
gurur	sevinç	**acı**	âb	**hırs**	kandırır	**anlay**	inandırır	**paranoyak**	sakındırmak	**anlamış**	öfkelenir

TABLE II. WORD LISTS USED TO BUILD *Lexicon2*

HAPPINESS	**FEAR**	**ANGER**	**SADNESS**	**DISGUST**	**SURPRISE**
mutluluk	korku	öfke	üzül	iğrenç	şaşır
sevgi	dehşet	kızgınlık	üzülür	ahlaksız	sanır
iyilik	umutsuzluk	umutsuzluk	affeder	gülünç	şaşırır
özlem	korkus	öfkes	pişma	acayip	uyandırır
duygu	çaresizlik	çaresizlik	inandırır	alaycı	hissettik
tutku	vahşet	pişmanlık	inanmaz	utanma	aptal
gurur	korkunç	hırs	üzüle	hissettik	anlamış
hayaller	zombi	suçluluk	anlamaz	yapmacık	hisset
şefkat	acı	acımasız	şüphelenir	aşağıla	korkar
heyecan	yalnızlık	dehşe	hatırlıyor	çıkarcı	hissediyor
sevgis	acımasızlık	kin	sanmak	çirk	söyley

The second stage of lexicon-based emotion detection covers the comparison of words in sentence and the words in emotion lexicon. For given sentence, this comparison operation simply generates a weight value for each emotion. In other words, an *emotion vector* of six values that represents the sentence where each value in vector refers to the weight of a specific emotion in sentence is constructed. *Emotion vectors* of sentences may be built in two different ways of comparison. Firstly, the exact matches to lexicon words may be considered as indicators as it is widely done in previous studies. Secondly, we propose to measure cosine distances of sentence words to lexicon words and employ them in *emotion vectors* of sentences. In our study, the *emotion vectors* are built in four different ways based on the comparison procedure and the strategy to obtain emotion values. These are *tf*, *MI-tf*, *max-similarity* and *average-similarity* vectors:

1) tf: In *tf* vectors, for each sentence, the words in sentence are compared to lexicon words one by one and the total number of matches to lexicon words of each emotion is summed up to build *emotion vector*. This approach may be accepted as the traditional way of building vectors. Simply in this approach, the high number of matches to the lexicon words of a specific emotion is considered as a strong indicator for the sentence to be assigned to the regarding emotion category. While *tf* method is applied, the sentences that includes no matches are omitted. In other words, if a sentence does not hold any matching word to *emotion lexicon*, the sentence is omitted from the experiments due to lack of evidence to classify it to one of the emotion categories.

2) MI-tf: In *MI-tf* vectors, matches to lexicon words are obtained for each sentence as in *tf* vectors. But in this approach, MI values given in [33] and [34] for matching words are summed up to obtain emotion values. In [34], it is mentioned that *MI* values, similar to *Lexicon3* words, are obtained from TREMO dataset by executing a set of complex operations. Since both lexicon words and their weights are obtained from TREMO dataset, the highest classification performance is expected to be observed when *Lexicon3* with *MI-tf* vectors is employed to classify the sentences in same data set due to the biased structure of the setting. In our experiments, we accept that this biased setting (*Lexicon3*, *MI-tf* vectors, data set: TREMO) is to produce the highest performance to be reached by proposed settings.

3) max-similarity: In *max-similarity* method, as an alternative to simple string matching, the cosine similarity of each word in sentence to each word in lexicon is calculated employing CBOW vectors obtained from Wikipedia data set. For each emotion, the most similar (the closest) lexicon word of regarding emotion is determined for each word in sentence. This similarity value is recorded as maximum similarity of the word. Following, for each sentence, the maximum similarity values are averaged to obtain the value of regarding emotion in *emotion vector*.

4) average-similarity: In *average-similarity*, similar to *max-similarity*, the cosine similarity to each lexicon word of a specific emotion is calculated for each word in sentence. For each emotion, the average of these similarity values is assigned as emotion value to the regarding emotion in *emotion vector* of sentence.

The last stage of emotion detection covers unsupervised and/or supervised labeling of sentences based on their *emotion vectors*. In supervised learning approach, emotion detection is accepted to be a classification task where given sentence is to be assigned to one of the six emotions. The emotion vectors are given as inputs to different classifiers. We split the data set in two as training and testing sets and applied 5-fold cross validation. In order to compare the performances of different classifiers in emotion detection, we employed Naïve Bayes (NB), Bayes Network (BN), Sequential Minimal Optimization (SMO), Random Forest (RF) and decision tree (J48) methods in Weka [36]. On the other hand, in unsupervised labelling, the sentence is classified to the emotion category that holds the highest emotion value in the vector. In addition, in case where there exists two equal highest emotion values in *emotion vector* of sentence, regarding sentence is accepted to be classified to both emotions and if one of them is the true category the result is accepted to be a hit (true positive).

IV. Experimental Results

In this study, two data resources are employed in the experiments. The first is TREMO dataset [34] that involves emotion labeled sentences. It was compiled in [34] where 5000 participants were asked to share their memories and experiences as text for six emotion categories. 27350 documents were collected from the participants. Due to the large number of participants and the inputs in the form of text, a verification process has been implemented. Each document was presented to three to five users, and the emotion category of the document was decided by majority of unanimous votes of 48 volunteers. In our experiments, TREMO sentences that are shorter than three words are ignored. In Table III, statistics on TREMO data set that is employed in our experiments are given.

The second data resource, Wikipedia that contains 4184516 articles in Turkish, is utilized to construct word vectors/embeddings. CBOW vectors of length=100 are built both for sentence and lexicon words. If the regarding word is not observed in Wikipedia, it is ignored in the experiments. Both data resources are subjected to a set of preprocessing operations to obtain the computable inputs to the experiments. Briefly, preprocessing covers the removal of punctuation marks, numerical characters, extra spaces and non-Turkish characters. Within preprocessing, the text is also subjected to Porter stemmer [37] and stop words are filtered.

We employed well-known accuracy (A), true positive rate (TPR) and F1 metrics in performance evaluation respectively in unsupervised and supervised experiments. A, TPR and F1 are given as

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

where TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives, respectively.

Table IV gives the statistics on *emotion vector* datasets in our experiment. For example, data set 1 refers to the set that is obtained by *tf* method and *Lexicon1*. Briefly, the words in *Lexicon1* is compared with the sentences in TREMO (a total of 20623 sentences) and in 4474 sentences the emotion vectors are obtained by at least one match to lexicon words. The data set 3 and 4 are the sets built by *Lexicon 3*. Though both the lexicon has been built from TREMO itself, in the experiments it is observed that there exists sentences in TREMO that do not contain any words of *Lexicon3*. We believe that this may be due to two reasons. The first is that due to the change in surface forms of words when different stemmers are utilized, the words may not match. The second is that following the retrieval of lexicon words from TREMO, the lexicon word list was subjected to improvement operations in [33]. These improvements may include the addition of new sentences.

In Table V, the accuracy values per emotion category are given when unsupervised approaches are followed. The bold values in each column refer to the top-most two accuracy values for regarding lexicon and method tuple; the last column per each method shows the average accuracy for given emotion in Table V. Considering Table V, following may be inferred:

1) Average accuracy values reveal that unsupervised lexicon-based approaches perform better in *disgust* and *surprise* emotions.
2) *Lexicon1* and *Lexicon2* continously succeed in emotion *surprise* regardless of method. On the other hand *Lexicon3* does not provide such consistent success for any emotions.
3) As the top-most performance values are examined *Lexicon2* generates higher scores compared to *Lexicon1* (except *max-similarity* method) as expected.
4) *Lexicon1* and *Lexicon2* dramatically fail in emotions *fear* and *anger*. On the other hand, *Lexicon3* provides acceptable accuracy results for these emotions, such that as average-similarity method is applied, *Lexicon3* provides its top-most scores.
5) *tf* method may be accepted in *happy*, *surprise* and *disgust* sentences (accuracy range [0,683- 0,976]) for all lexicons.
6) *max-similarity* method provides succesfull classification in emotion *disgust* (accuracy range [0,793-0,928]) and fails in *sadness* (accuracy range [0,375-0,493]) for all lexicons.
7) In average similarity method, there exists no emotion that all lexicons succeed.

In Table VI, average accuracy results per method and lexicon duo are given. To exemplify, when *Lexicon1* is used in *tf* method, the overall accuracy is observed to be $0,585$. It is examined that the highest accuracy values $0,740$ and $0,794$ belong to *tf* method with *Lexicon2* and *MI-tf* with *Lexicon3*. Considering that *Lexicon3* is actually built up utilizing the TREMO itself, such a high accuracy value is not surprising. Besides *tf* method and *Lexicon2* duo provides almost similar performance. The disadvantage of *Lexicon2* is that via it involves only 66 lexicon words, the number of sentences to be classified is limited compared to *Lexicon3*. On the other hand, as the effort required to build up the lexicons and to obtain lexicon word weights are compared, it can be stated that *Lexicon2* provides promising results and it is worth to generate a larger lexicon set by the proposed method as a further work.

In Table VII, performance values in supervised experiments

TABLE III. TREMO DATASET USED IN EXPERIMENTS

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total
Number of sentence	3877	2634	3206	4135	4048	2723	20623
Average sentence length (Number of words)	4,99	4,80	5,10	4,78	4,93	5,37	4,98

TABLE IV. DATASETS

2*Data set	2*Lexicon	2*Method	Data set size (Number of sentences)
1	*Lexicon1*	*tf*	4474
2	*Lexicon2*		3583
3	*Lexicon3*		15180
4	*Lexicon1*	*average-similarity*	20623
5	*Lexicon2*		
6	*Lexicon3*		
7	*Lexicon1*	*max-similarity*	20623
8	*Lexicon2*		
9	*Lexicon3*		
10	*Lexicon3*	*MI-tf*	15180

TABLE V. ACCURACY RESULTS PER EMOTION CATEGORY - UNSUPERVISED EXPERIMENTS

	tf				*max-similarity*				*average-similarity*			
Emotions	*Lexicon1*	*Lexicon2*	*Lexicon3*	Average	*Lexicon1*	*Lexicon2*	*Lexicon3*	Average	*Lexicon1*	*Lexicon2*	*Lexicon3*	Average
Happiness	0,697	0,726	**0,813**	0,745	0,429	0,399	0,710	0,513	0,389	0,535	0,169	0,364
Fear	0,230	0,316	0,556	0,367	0,409	0,376	0,615	0,467	0,287	0,347	**0,723**	0,452
Anger	0,073	0,391	0,675	0,380	0,088	0,078	**0,869**	0,345	0,192	0,190	**0,802**	0,395
Sadness	**0,913**	**0,932**	0,432	0,759	0,487	0,375	0,493	0,452	0,536	0,496	0,355	0,462
Disgust	0,731	0,771	**0,747**	**0,750**	**0,731**	**0,928**	**0,793**	**0,817**	**0,679**	**0,745**	0,159	**0,528**
Surprise	**0,957**	**0,976**	0,683	**0,872**	**0,916**	**0,859**	0,445	**0,740**	**0,862**	**0,872**	0,087	**0,607**

TABLE VI. AVERAGE ACCURACY RESULTS - UNSUPERVISED EXPERIMENTS

	Lexicon1	*Lexicon2*	*Lexicon3*
tf	0,585	0,740	0,630
MI-tf	-	-	0,794
max-similarity	0,476	0,459	0,658
average-similarity	0,465	0,505	0,399

are presented. Similar to Table V, bold values in columns indicate top-most performance values for regarding method.

Examining the classification methods that give the highest *F1* values in Table VII, it is observed that in 4 of 10 sets RF gives the acceptable highest performance values. Though SMO method gives highest scores for 6 settings, it cannot be considered as a succeeding method due to the *F1* values lower than 0.5.

The two columns on right in Table VII indicate the average performance results of supervised learning methods. Average results indicate the following:

1) The first two data sets (1 and 2) that are compiled by *tf* method generate promising highest performance values in emotion classification in both average *F1* and *TPR* measures.
2) Though the data set (3) is actually constructed from TREMO set, it failed to classify the sentences in TREMO.
3) The data sets (4-9) that are built by *average-similarity* or *max-similarity* are examined to fail in emotion classification.

V. DISCUSSION AND CONCLUSION

In this paper, we aimed to build a hybrid approach that reduces the effort required in emotion detection by revealing the strengths of keyword and lexicon-based approaches. To this aim, we proposed the use of word embeddings in two main tasks of emotion detection process. Firstly, embeddings are employed in *emotion lexicon* construction task in order to decrease the human effort in labeling by reducing the number of arbitrary words. In this task, the list of words belonging to

TABLE VII. *TPR* AND *F1* RESULTS - SUPERVISED EXPERIMENTS

Data set	Classification Method	*TPR*	*F1*	Average *TPR*	Average *F1*
1 (Lexicon1+ tf)	BM	0,720	0,686	**0,725**	**0,689**
	NB	0,718	0,681		
	SMO	**0,728**	**0,692**		
	J48	0,727	0,690		
	RF	**0,731**	**0,697**		
2 (Lexicon2+ tf)	BM	0,723	0,687	**0,736**	**0,706**
	NB	0,722	0,694		
	SMO	0,744	0,716		
	J48	**0,745**	**0,717**		
	RF	**0,745**	**0,718**		
3 (Lexicon3+ tf)	BM	0,546	0,536	0,575	0,562
	NB	0,496	0,473		
	SMO	0,593	0,587		
	J48	**0,618**	**0,604**		
	RF	**0,623**	**0,610**		
4 (Lexicon1+ average-similarity)	BM	0,239	0,165	0,299	0,245
	NB	0,238	0,164		
	SMO	**0,425**	**0,382**		
	J48	**0,355**	**0,350**		
	RF	0,239	0,165		
5 (Lexicon2+ average-similarity)	BM	0,246	0,171	0,313	0,258
	NB	0,244	0,169		
	SMO	**0,451**	**0,409**		
	J48	**0,377**	**0,371**		
	RF	0,246	0,171		
6 (Lexicon3+ average-similarity)	BM	0,342	0,300	0,393	0,362
	NB	0,346	0,306		
	SMO	**0,511**	**0,490**		
	J48	**0,422**	**0,416**		
	RF	0,342	0,300		
7 (Lexicon1+ max-similarity)	BM	0,249	0,179	0,302	0,256
	NB	0,249	0,182		
	SMO	**0,420**	**0,399**		
	J48	**0,343**	**0,339**		
	RF	0,249	0,179		
8 (Lexicon2+ max-similarity)	BM	0,249	0,182	0,308	0,263
	NB	0,248	0,182		
	SMO	**0,436**	**0,414**		
	J48	**0,358**	**0,353**		
	RF	0,249	0,182		
9 (Lexicon3+ max-similarity)	BM	0,430	0,422	0,445	0,436
	NB	0,431	0,422		
	SMO	**0,497**	**0,482**		
	J48	**0,436**	**0,432**		
	RF	0,430	0,422		
10 (Lexicon3+ MI-tf)	BM	0,613	0,616	0,564	0,571
	NB	0,433	0,462		
	SMO	0,462	0,487		
	J48	**0,656**	**0,644**		
	RF	**0,654**	**0,644**		

an emotion category is determined by measuring the vector-based similarity to predetermined keywords. The second is that word embeddings are used while sentences are compared to lexicon words in order to be labelled to either one of 6 emotion categories. In this task, sentences are represented by *emotion vectors* and four alternative approaches to build these vectors are presented. The distance between *emotion vectors* and lexicon word embeddings are measured in order to decide the emotion label of the regarding sentence.

The performance of the proposed approaches are examined both in supervised and unsupervised emotion detection experiments. In the experiments, the success of presented lexicons are compared to an existing lexicon that is accepted to be the base set. It is shown that the emotion detection scores vary for different emotions for all lexicons and no lexicons perform significantly better than others in emotion detection task.

Considering the set of four alternative approaches to build *emotion vectors*, the proposed vectors are evaluated relative to the base emotion vector that is built employing preexisting weighting scheme. It is observed in both supervised and unsupervised experiments that though performance scores are lower than expectations, the scores of proposed approaches are promising compared to base emotion vectors.

Based on the experimental results, it is examined that the use of word embeddings in lexicon construction is encouraging such that it is worth to enlarge regarding lexicons as a future work. Beside, the use of word embedding similarities in emotion identification stage; in other words, building emotion vectors based on cosine similarity; did not succeed compared to exact match strategy. As a future work, we plan to enhance our lexicons where word vector similarity is employed to determine lexicon words. In addition, we will run experiments with succeeding methods on different datasets.

References

[1] Garcia J, Penichet VMR, Lozano MD. Emotion detection: a technology review. In: Proceedings of the XVIII International Conference on Human Computer Interaction (Interacción ’17), 2017; Association for Computing Machinery, New York, NY, USA, Article 8, 1–8. DOI:https://doi.org/10.1145/3123818.3123852

[2] Ekman P. An argument for basic emotions. Cognition and Emotion, 1992; 6(3/4), pp. 169-200.

[3] Plutchik R. A general psychoevolutionary theory of emotion. In: Emotion: Theory, Research and Experience, 2019; New York: Elsevier Inc., pp. 1: 3-33.

[4] Mikolov T et al. Distributed Representations of Words and Phrases and their Compositionality. In: NIPS’13, Proceedings of the 26th International Conference on Neural Information Processing Systems; 2013, pp 3111–3119

[5] Sailunaz K, Dhaliwal M, Rokne J, Alhajj R. Emotion detection from text and speech: a survey. Social Network Analysis and Mining 2018; 8, 1-26.

[6] Chopade CR. Text based emotion recognition: a survey. International Journal Science Research 2015; 4(6), pp 409-414.

[7] Canales L, Martinez-Barco P. Emotion detection from text: a survey. In: Processing in the 5th Information Systems Research Working Days (JISIC 2014); 2014. pp. 37-43

[8] Binali H, Wu C, Potdar V. Computational approaches for emotion detection in text. In: 4th IEEE International Conference on Digital Ecosystems and Technologies; 2010. pp. 172-177

[9] Al-Saaqa S, Abdel-Nabi H, Awajan A. A Survey of Textual Emotion Detection. In 2018 8th International Conference on Computer Science and Information Technology (CSIT); 2018. pp. 136-142, doi: 10.1109/CSIT.2018.8486405

[10] Alswaidan N, Menai MEB. A survey of state-of-the-art approaches for emotion recognition in text. Knowl Inf Syst 2020; 62:2937–2987, https://doi.org/10.1007

[11] Strapparava C, Mihalcea R. Learning to identify emotions in text. In: Proceedings of the 2008 ACM symposium on Applied computing – SAC ’08; 2008. pp 1556–1560, New York, New York, USA. ACM Press.

[12] Fellbaum C. WordNet: An Electronic Lexical Database. London, UK: MIT Press, 1998. Sailunaz K, Dhaliwal M, Rokne J, Alhajj R. Emotion detection from text and speech: a survey. Social Network Analysis Mining, 8 (1), 2018.

[13] Poria S, Gelbukh A, Hussain A, Das D, Bandyopadhyay S. Enhanced SenticNet with Affective Labels for Concept-based Opinion Mining. IEEE Intelligent Systems 2013; vol. 28, issue 2, pp. 31–38, doi:10.1109/MIS.2013.4.

[14] Poria S, Gelbukh A, Das D, Bandyopadhyay S. Fuzzy Clustering for Semi-Supervised Learning—Case study: Construction of an Emotion Lexicon. In: MICAI 2012, Advances in Computational Intelligence, Lecture Notes in Artificial Intelligence; 2012. pp. 73–86.

[15] Poria S, Gelbukh A, Cambria E, Yang P, Hussain A, Durrani T. Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis. In: IEEE 11th International Conference on Signal Processing (ICSP); 2012. Beijing, Vol.2, pp.1251–1255; doi: 10.1109/ICoSP.2012.6491803.

[16] Poria S, Gelbukh A, Cambria E, Das D, Bandyopadhyay S. Enriching SenticNet Polarity Scores through Semi-Supervised Fuzzy Clustering. In: Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, IEEE 12th International Conference on Data Mining Workshops (ICDMW); 2012. Brussels, Belgium. IEEE CS Press, pp. 709–716, doi: 10.1109/ICDMW.2012.142.

[17] Mohammad S, Turney P. Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 2013; 29(3): 436-465, Wiley Blackwell Publishing Ltd.

[18] Mohammad S, Turney P. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In: Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text; 2010. LA, California.

[19] Staiano J, Guerini M. Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2014. Association for Computational Linguistics, Baltimore, Maryland, https://doi.org/10.3115/v1/P14-2070

[20] Gupta U, Chatterjee A, Srikanth R, Agrawal P. A sentiment-and-semantics-based approach for emotion detection in textual conversations. In: Neu-IR Workshop on Neural Information Retrieval, SIGIR; 2017. ACM, arXiv preprint arXiv:1707.06996

[21] Desmet B, Hoste V. Emotion detection in suicide notes. Expert Syst Appl 2013; 40(16):6351–6358

[22] Hajar M. Using youtube comments for text-based emotion recognition. Procedia Comput Sci 2016; 83:292–299

[23] Agrawal A. Unsupervised emotion detection from text using semantic and syntactic relations. In: Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology; 2012. pp346–353

[24] Dehkharghani R, Saygın Y, Yanikoglu B, Oflazer K. Sentiturknet: a Turkish polarity lexicon for sentiment analysis. Language Resources and Evaluation 2015; 50:667-685.

[25] Bilgin O, Cetinoglu O, Oflazer K. Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology 2004; 7(1-2):163-172.

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

[27] Kaynar O, Gormez Y, Yildiz M, Albayrak A. Sentiment analysis with machine learning techniques. In: International Artificial Intelligence and Data Processing Symposium; Malatya, Turkey; 2016. pp. 80-86.

[28] Ucan A, Naderalvojoud B, Sezer EA, Sever H. SentiWordNet for new language: automatic translation approach. In: 12th International

Conference on Signal-Image Technology and Internet-Based Systems; Naples, Italy; 2016. pp.308-315.

[29] Ersahin B, Aktas O, Kilinc D, Ersahin M. A hybrid sentiment analysis method for Turkish. Turkish Journal of Electrical Engineering and Computer Sciences 2019; 1780-1793.

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

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

[32] Scherer KR, Wallbott HG. The ISEAR questionnaire and codebook. 1997, Geneva Emotion Research Group.

[33] Tocoglu MA, Alpkocak A. Lexicon-based emotion analysis in Turkish. Turkish Journal of Electrical Engineering and Computer Sciences 2019; 27(2):1213-1227.

[34] Tocoglu MA, Alpkocak A. TREMO: A dataset for emotion analysis in Turkish. Journal of Information Science 2018.

[35] Tocoglu MA, Ozturkmenoglu O, Alpkocak A. Emotion Analysis From Turkish Tweets Using Deep Neural Networks. In IEEE Access 2019; 7:183061-183069, doi: 10.1109/ACCESS.2019.2960113.

[36] Witten IH. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In: Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems; 1999. pp192-196.

[37] Porter MF. An algorithm for suffix stripping. Program: Electronic Library and Information Systems 1980. 40(3), 211-218.