IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 1, 2019, accepted December 12, 2019, date of publication December 16, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960113

Emotion Analysis From Turkish Tweets Using Deep Neural Networks

MANSUR ALP TOCOGLU[1], OKAN OZTURKMENOGLU[2], AND ADIL ALPKOCAK[2], (Member, IEEE)

[1]Department of Software Engineering, Manisa Celal Bayar University, 45400 Turgutlu, Turkey
[2]Department of Computer Engineering, Dokuz Eylül University, 35160 İzmir, Turkey

Corresponding author: Adil Alpkocak (alpkocak@ceng.deu.edu.tr)

ABSTRACT Text data analysis of social media is becoming more and more important since it includes the most recent information on what people think about. Likewise, emotion is one of the most valuable parts of human communication, emotion analysis is a type of information extraction process which identifies the emotional states of a given text. In this study, we investigated the performance of deep neural networks on emotion analysis from Turkish tweets. For this, we examined three different deep learning architectures including artificial neural network (ANN), convolutional neural network (CNN) and recurrent neural network (RNN) with long short-term memory (LSTM). Besides, we curated a dataset of Turkish tweets and annotated each tweet automatically for six emotion categories using a lexicon-based approach. For the evaluation, we conducted a set of experiments for each architecture. The results showed that the lexicon-based automatic annotation of tweets is valid. Secondly, ANN produced the worst result as expected, and CNN resulted in the highest score of 0.74 in terms of accuracy measure. Experiments also showed that our proposed approach for emotion analysis of tweets in Turkish performs better than state-of-the-art in this topic.

INDEX TERMS Emotion analysis; Twitter, deep learning, Turkish text analysis, text mining, machine learning, information extraction.

I. INTRODUCTION

Nowadays, social media plays an important role in our daily life. Every day, millions of people share their thoughts, feelings, experiences through microblogging services such as Twitter, Facebook, Instagram and so on. As people like to share their ideas and interests, social media becomes a more important source of information. Moreover, microblogging services include the most recent information about what is going on in the world. All these features make social media an indispensable resource for understanding the community. However, working with social media data includes some difficulties in its raw form, since it is full of misspelled words, strange abbreviations and weird jargon that are not used in everyday spoken language. Despite all, social media includes invaluable clues of human daily life.

Emotion is one of the most valuable information for human communication because they are undisputed parts of the lives of humans as they are a part of humans from the beginning of this world. In other words, they are innate. In the literature, some studies proposed a taxonomy of basic emotions [1], [2]. In one of these studies, Ekman proposed a model of six basic emotions which are *joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise* [1]. Even though this model includes concrete boundaries between the emotion categories, generally, it is not that trivial to define clear boundaries between them [3]. Humans inherently understand the emotions in human behavior. However, understanding of emotions from a given text is still an open issue for computer science, where emotion analysis is defined as a process of information extraction identifying emotional states of a given text. Furthermore, detecting emotions in microblogs and social media posts is a popular research topic in many different application domains.

Emotion analysis using supervised machine learning algorithms is popular and provides better results compared to unsupervised learning algorithms [4]. However, supervised approaches require a labeled training set, which is time and labor-intensive operation. In literature, datasets were mostly annotated automatically based on either using a set of keywords in hashtags [4]–[7] or emoticons, for emotion

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

analysis [8], [9] and sentiment analysis [10]–[12]. In these studies, they mostly focused on hashtags rather than considering the whole tweets. Surely, use of hashtags intensifies or highlights the meaning. However, focusing only hashtags may cause to miss some possible emotive information in a sentence. Hence, we intended to use a larger lexicon for annotation considering the whole tweet. Furthermore, there are plenty publicly available datasets for English, however, to the best of our knowledge, there is no publicly available dataset for Turkish. Additionally, there are some studies about supervised emotion analysis of tweets in Turkish [5], [6], but no one using deep neural networks on a large tweet dataset. Consequently, there is a need for a performance comparison of different deep learning architectures for emotion analysis of tweets in Turkish.

Contribution of this paper is two-fold: First, it presents a deep learning-based emotion analysis approach and provides a comparison of different architectures, such as a convolutional neural network (CNN), recurrent neural network (RNN) with long short-term memory (LSTM) and artificial neural network (ANN). The second contribution is the curation of a new dataset, which we named as Turkish Twitter emotion dataset (TURTED), for the use of Turkish emotion analysis. It is open to the public for academic use (available at *http://demir.ceng.deu.edu.tr/turted*). To the best of our knowledge, this is the first automatically annotated dataset of tweets in Turkish. It includes more than 195K tweets annotated in six emotional states for *happiness*, *sadness*, *fear*, *anger*, *disgust*, and *surprise*, using a lexicon-based automatic annotation method using the Turkish emotion lexicon (TEL) [13].

The rest of the paper is organized as follows: The next section presents a comprehensive introduction of recent studies done in the literature on this topic. Then, section three gives technical details about the material and the methodology we used in both the database curation process and emotion analysis using a deep learning approach. section four presents the experiments we conducted for this study and provides a discussion on the findings of comparing different approaches for emotion analysis in Turkish. Section five concludes the paper and provides a projection for further studies on this topic.

II. RELATED WORK

Twitter plays an important role in providing raw data to be used in sentiment and emotion analysis. In literature, there are many studies about sentiment [14]–[17] and emotion analysis [18]–[20] on Twitter data in many languages including Turkish [6]. However, most studies in Turkish text deals with sentiment analysis rather than emotion analysis. So, in this section, we provide the state-of-the-art in sentiment and emotion analysis of Twitter data in both English and Turkish.

In the literature, some studies dealing with sentiment analysis of Twitter data in Turkish. Çoban *et al.*, [21] focused on analyzing sentiment extraction from social media sources. So, they first collected a dataset composed of 14,777 tweets in Turkish. They categorized the tweets in two sentiment categories, positive and negative by using basic sentiment icons, and obtained 66.06% accuracy using a Multinomial Naïve Bayes classifier. In another study [22], Akgül et al. developed a software application to fetch positive, negative or neutral labeled Turkish tweets from Twitter data using a given lexicon. They proposed two approaches using lexicon and *n*-grams. Then, they evaluated the system in terms of F-measure scores. The results showed that the lexicon-based method slightly outperformed the *n*-gram method where the F-measure scores were 70% and 69%, respectively.

The number of studies about emotion analysis is relatively less than sentiment analysis. The scarcity of emotion analysis in Turkish is not limited only to Twitter data, it is the case for all text data. In the study [19], Boynukalin used two datasets, which were Turkish translation of ISEAR [23] dataset and Turkish fairy tales, to do emotion analysis, for four categories *joy*, *sadness*, *anger*, and *fear*. Tocoglu and Alpkocak [24] generated an emotive dataset the TREMO to be used in emotion analysis in Turkish for emotion categories *happiness*, *fear*, *anger*, *sadness*, *disgust*, and *surprise*. In another study [13], Tocoglu and Alpkocak created the TEL for the use of lexicon-based emotion analysis in Turkish by using TREMO dataset. Demirci [6] focused on extracting the emotion of Twitter data in Turkish, and collected tweets for the six emotions, *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* using the Twitter hashtags for each emotion category. Demirci defined hashtags containing the derivatives of each emotive word, and as a result, 6,000 tweets, 1,000 tweets for each emotion, were collected in total. They investigated the effects of different machine learning algorithms of naive Bayes, complement naive Bayes, support vector machine, and *k*-NN classifiers. It was reported that the support vector machine outperformed the others by achieving 69.92% of accuracy. However, the dataset is not publicly available and not large enough to be used in deep learning architectures. In another study [25], Ileri focused on predicting emotions of users on a subject on Twitter by a different user-centric approach using relationship information of the users. To do so, they constructed a dataset, which is composed of 7,200 tweets in total, 1,200 tweets for each category for six emotion categories.

In the literature, there are many more studies in emotion analysis of tweets dealing with English rather than Turkish. Jabreel and Moreno [18] dealt with the emotion classification of Twitter for eleven emotion categories. These are Plutchik's eight categories [26] (i.e., *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*) and three additional categories for love, optimism, and pessimism. The authors proposed a novel method where the multi-label classification problem was transformed into a binary classification problem first and then solved by using deep learning models. They used a dataset composed of 10,983 samples [27] for training and testing. The proposed system outperformed the state-of-the-art systems with an accuracy score of 0.59. Bandhakavi *et al.*, [20] focused on emotion extraction using

the knowledge of domain-specific lexicons (DSELs) and general-purpose emotion lexicons (GPELs). They extracted features using lexicons by using their unigram mixture model (UMM) and they compared emotion classification results on four benchmark datasets, SemEval-2007 [28], Twitter dataset [29], blog dataset [30], and ISEAR [23]. As a result, the use of DSELs performed better than GPELs. Sailunaz and Alhajj [31] generated a personalized recommendation system for the Twitter activities of users. To do so, first, they collected 7,246 tweets and replies in total, then gathered 3,607 user information related to collected tweets and replies. The collected raw dataset was manually annotated for emotions of *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *neutral*, and their polarity as *positive*, *negative* or *neutral*. Then the authors' calculated the influence score of users by using sentiment score and emotion score to generate general and personalized recommendations for users based on their Twitter activities.

In another study [32], Qadir and Riloff used a bootstrapping algorithm to identify emotion hashtags automatically for each emotion category of *affection*, *anger*, *fear*, *joy*, and *sadness*. The algorithm starts with training classifiers by using a small number of labeled hashtags to identify new emotion hashtags. This process continues iteratively until generating a learned hashtag list which improves emotion classification performance. In the study [7], Mohammad and Bravo-Marquez generated a Twitter dataset and then annotated for the emotion categories of *anger*, *fear*, *joy*, and *sadness*. To improve annotation consistency, the authors used the Best-Worst Scaling technique, and created a regression analysis to determine useful features to be used for emotion intensity classification. Mohammad [4] focused on generating Twitter Emotion Corpus (TEC) containing 21,051 tweets by using emotion-word hashtags for Ekman's six emotion categories. Next, the author conducted experiments to analyze whether the self-labeled TEC dataset matches with annotation process done by trained annotators. Consequently, it is proved that the self-labeled hashtag annotations are consistent. The author created a word-emotion association lexicon using the corresponding Twitter dataset to prove that it provides higher emotion classification results than lexicons generated manually. In another study [31], the authors focused on detecting sentiment and emotion expressions from tweets and generated a personalized recommendation system for the users' Twitter activity. Hasan *et al.*, [33] generated a supervised learning system to extract emotion automatically from live streams of text messages for real time emotion tracking.

III. MATERIAL AND METHOD

In this section, we present a brief introduction about the Turkish language and then give the details of the dataset curation process. Next, we discussed the details of the method we used in the evaluation phase.

A. TURKISH LANGUAGE

Turkish is a member of the Oghuz group of the Turkic languages, which belongs to the Altaic branch of the Ural-Altaic language family. Turkish uses a Latin alphabet consisting of 29 letters (**a**, b, c, ç, d, **e**, f, g, ğ, h, **ı**, **i**, j, k, l, m, n, **o**, **ö**, p, r, s, ş, t, **u**, **ü** , v, y, z), of which 21 are consonants and 8 are vowels (written in boldface). It is a highly agglutinative language similar to Finnish and Hungarian. It has very productive inflectional and derivational processes and so it is a morphologically complex language.

In Turkish, the meaning of a words can be changed when a suffix is added. For example, the word ''göz'' (i.e., eye) may take a suffix ''lük'', so it becomes ''gözlük'' (i.e., glasses). Then, if it takes another suffix ''çü'', it becomes ''gözlükçü'' (i.e., optician). Furthermore, it may take another suffix ''lük'', so it becomes ''gözlükçülük'' (i.e., the profession of an optician). As shown in the example, each suffix may create a new word having a totally new meaning.

Building up of words through suffixes allow a complex concept might be expressed in a single word in Turkish. For example; ''Gerçekleştirilemeyenlerdir'' is a possible one-word sentence starting from the adjective ''gerçek'' (i.e., real). It is translatable in English as ''Those are the things which could not be put into practice [realised]''.

Turkish has a set of rules when adding suffixes to a word. One of them is known as vowel harmony rule regulating the change of the last consonant of a word stem as some new suffixes are added to it. This is a phonological process to ensure a smooth flow and forcing the least amount of oral movement as possible when a series of suffixes are added to the stem word [34]. For example, with the suffix ''a'', ''kita**p**'' (i.e., book) becomes ''kita**b**a'' (i.e., towards the book) while with the suffix ''da'', the root does not change and becomes ''kita**p**ta'' (i.e., ''in the book''). In this example, note the transformation of ''da'' to ''ta'' due to the last letter ''p'' in word stem. Another typical example to this type of rules is vowel drop, which forces some vowels in the word body to drop out in some word and suffix combinations. For example, with the suffix ''um,'' the boldface letter **u**drops in ''oğ**u**l'' (i.e., son) and becomes ''oğlum'' (i.e., my son). Turkish language is mostly regular but all these makes Turkish language hard to analyze lexically and morphologically.

Another distinctive characteristic of Turkish language is being a free-word order language [35]. Turkish has no noun classes or grammatical gender and has a basic word order of subject–object–verb, but can be changed. When the order of words changes in a sentence, the new generated sentences have mostly same meaning with a slightly difference in emphasis. For example, considering the sentence ''Pazartesi günü uçakla Ankara'ya gittim'' (i.e., I went to Ankara by plane on Monday), the followings can be possible orders: ''Ankara'ya Pazartesi günü uçakla gittim'', ''Uçakla Pazartesi günü Ankara'ya gittim'', ''Ankara'ya uçakla Pazartesi günü gittim'' and so on. They are all in different order with the

TABLE 1. Samples keywords of TEL for each emotion category, where English translations are given in italic and parentheses.

Emotion Category	TEL Keywords
Fear	kaybol (get lost), trafik kaza (traffic accident), asansör (elevator), sigara (cigaret), deprem oldu (earthquake happened), rüya (dream), yalnızlık (loneliness), köpek (dog), örümcek (spider), film izle (watch the film)
Happy	oyna (play), eğlen (have fun), geçir (undergo), dedem (my grandfather), müzik dinlemek, (listen to music), sevindi (delighted), şampiyon (champion), sınav (exam) alışveriş (shopping), sigara (cigaret)
Disgust	sümük (snot), iğrenç (disgusting), geçir (undergo), içme (drinking), yalnız (alone), babam (my father), bulaşık (dishes), sigara (cigaret) ölü (dead), sakız (chewing gum)
Anger	kardeş (sibling), yalan söyleme (lying), bencil (selfish), söyleme (saying), sorma (asking), salak (fool), ölme (death), insan nefret (human hate), sınav (exam), ağız (mouth)
Sadness	kuşu (bird), içim (inner), ağla (weep), canım (sweetheart), yapma (making), doğum (birth), ayrıl (leave), ölmüş (deceased), üzer (sad), yanım (my side)
Surprise	şaşırt (surprise), öğren (learn), hayret (astonishment), köpek (dog), geçir (undergo), şaşırma (astonishment), alınca (taken by the), arkadaş (friend), bekleme (waiting), mutlu olmuş (was happy)

TABLE 2. Sample tweets and their annotations, where English translations are given in italic and parentheses.

Tweet Id / Tweet text (*English*)	Category (TotalScore)	TEL Keyword (*English*) (MI value)
1114992356992987147 / Küçükken benim de iki tane ördeğim vardi kardesimle benim ismimizi koymuştuk.sonra ananemle **dedem** onlari kesip **yedi** (*When I was little, I had two ducks. My brother and I named him after me. Then **my** grandmother and **grandfather** kill them and **ate** them.*)	**Sadness** (0. 014451)	dedem (my grandfather) (0.014451)
		yedi (*ate*) (0.002897)
	Disgust (0.004242)	dedem (*my grandfather*) (0.001345)
1113545562308120577 / Canım oğlum sen sırf pubg **oyna** diye annen sana en iyi oyun bilgisayarı alacak. Ama takdir getirirsen. Babanı hallederim sen merak etme. (*My dear son, your mother will get you the best gaming computer just because you **play** Pubg. But if you bring me a certificate of appreciation. Don't worry, I'll take care of your father.*)	**Happy** (0.001724)	oyna (*play*) (0.001724)
1114126715461296128 / Otobüsteki çocuk az daha zorlarsa **sümük** yerine beynini çıkaracak (*The kid on the bus will pull out his brain instead of **snot** if he tries harder.*)	**Disgust** (0.002301)	sümük (*snot*) (0.002301)

same meaning, which emphasizes the word closer to the verb (''gittim'', i.e., I went) at the end.

B. CURATION OF DATASET

The dataset we curated, which we named as Turkish Twitter emotion dataset, contains Turkish tweets from Twitter which is commonly used social networking and microblogging service. In Twitter, registered users can read and post messages about every conceivable subject that is named as tweets. Each tweet has a limit of 280 characters and also its context can be solecistic, from daily lives to the developments in nature and society. So, people express their emotions about these developments around them in tweets. The datasets created with tweets may contain a lack of context, spelling errors on purpose or not, slang and repeating characters.

To annotate tweets, we have intensively used TEL, which consists of 7,235 keywords, in total, for six emotion categories of *fear*, *happy*, *disgust*, *anger*, *sadness*, *surprise* [36]. Table 1 shows sample keywords selected from TEL for each emotion category, where English translations are presented in italic. We have chosen 879 unique keywords in total for six emotion categories. Then, we used each keyword in data curation. Additionally, TEL includes Mutual Information (MI) value for each keyword for each emotion category as a weight.

We curated dataset in two steps: data gathering and annotation. In the data gathering step, we collected 205,278 tweets via Tweepy which is a Python library [37] providing the connection to Twitter API. Some tweets contain multi-keywords of different emotion categories, where one tweet is stored in the dataset more than once for different emotions. In the dataset, some tweets include keywords belong to more than one emotion category. Table 2 shows some sample tweets containing keywords belong to more than one emotion category.

We set a filter in Tweepy for selecting tweets in Turkish and to not include retweets. Later, we discarded tweets containing the tokens ''RT @'' and which stands for inline retweets. The data we collect includes tweets starting from April 4, 2019, to April 9, 2019, which is just a few days after the local government elections held in Turkey. This is important because the contents of tweets may be related to the agenda of that time period.

For annotation, we performed a lexicon-based automatic annotation process using TEL [36], which includes keywords lemmatized by TurkLemma [38]. Then, we annotated each tweet with a category name of a TEL keyword contains in that tweet. In TEL, an emotive keyword may appear more than one emotion category. Inherently, a tweet might be considered to annotate more than one emotion category with a single keyword. However, we labeled a tweet with a single emotion category. To do this, we simply choose the category whose *EScore* value is the maximum. Thus, we summed the MI values of each TEL keyword for a corresponding emotion category. Then, we selected the category whose value is the highest as the category label of the tweet. More formally, let us assume that $Val\ (k, j)$ is a function giving MI value of TEL keyword k, for emotion category of j, as follows:

$$Val\ (k, j) = \begin{cases} w, & if\ k \in List\ (e_j) \\ 0, & otherwise \end{cases} \quad (1)$$

TABLE 3. Emotion category distribution of the dataset TURTED.

Emotion	# of TEL Keywords	Unique Users	Total Tweets	Avg. length	Total token count	Avg. token length
Fear	136	23,652	33,144	134.65	583,305	6.62
Happy	107	20,501	27,735	141.99	516,959	6.59
Disgust	165	23,103	32,275	127.07	554,944	6.37
Anger	185	31,079	44,323	139.81	822,066	6.52
Sadness	159	24,225	34,229	122.74	567,333	6.38
Surprise	127	17,736	23,739	142.00	447,718	6.51
Total	**879**	**140,296**	**195,445**	-	**3,492,325**	-

where $List\left(e_j\right)$ denotes the TEL keyword k of belonging to the j^{th} emotion category and the value of w is the MI value. Additionally, for a given tweet of t, and emotion category of j, the total emotion score, *EScore*, is defined as follows:

$$EScore\,(t,j) = \sum_{k \in List\,(e_j)}^{\|t\|} Val\,(k,j) \tag{2}$$

where $\|t\|$ indicates length of a tweet in number of tokens. Then, the category having the maximum *EScore* is selected as classification result, as follows:

$$Category \leftarrow \underset{i=1,n}{arg\,max}\ (EScore(t,i)) \tag{3}$$

where n is the number of category and assuming that tweet t contains keywords belonging to different emotion categories. This approach considers the keywords only in TEL which may be limited, because of words that are closely related to the keywords in TEL but not included in it. These keywords might be included by using an external knowledge.

Table 2 presents some example tweets, TEL keywords, and their emotion categories. Table 3 shows the count of TEL keywords, users, tweets, and some statistics for the dataset. The first column has an emotion category and the second column shows the number of TEL keywords in each emotion category. When curating the dataset, we filtered tweets based on whether they contain these keywords. The third column has the count of unique users that post the tweets gathered for an emotion category. The next column presents the total number of tweets in TURTED for each emotion category. The fifth column indicates the average character length of tweets. The next column is the total number of tokens contained in each category. The last column shows the average character length of tokens.

C. PREPROCESSING

The goal of the preprocessing phase is to prepare the collected Twitter raw data to use in the experimental procedures. First, we focused on eliminating terms starting with ''http'', which indicates links to some web sites and frequently used in tweets. Second, we deleted the punctuation characters, the extra spaces, and all numeric characters. Then, we preprocessed the dataset using two different stemmer approaches, which are fixed length (F5, i.e., taking the first five characters of a term) [34] and Snowball stemmer (SS) [39]. At last, we removed stop-words for the Turkish language provided in Python natural language toolkit (NLTK).

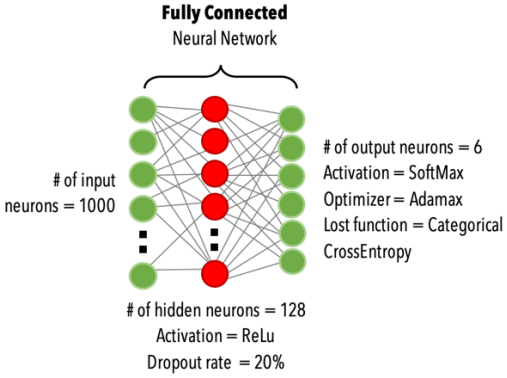

FIGURE 1. Architecture of the fully connected neural network used in this study.

D. EMOTION ANALYSIS

In the literature, deep learning methods are frequently used in machine learning tasks such as image and voice classification problems [40], [41]. However, in recent studies, they also provide higher results in natural language processing tasks compared to traditional sparse and linear models [42]. In this study, we compared three different deep neural network architectures, these are ANN, CNN, and LSTM.

1) ARTIFICIAL NEURAL NETWORK

Fig. 1 shows the architecture of the sequential dense artificial neural network we used in the experimentation phase. The architecture is composed of three layers which are input, hidden, and output layers. The input layer is composed of 1000 input neurons whereas the other two layers contain 128 and 6 neurons, respectively. We chose a network dropout rate as 20% and the activation functions of the neurons in the hidden and output layers as ReLu and SoftMax functions. Besides, we used Adamax and Categorical *CrossEntropy* functions for the lost and optimizer parameters of the network.

2) CONVOLUTIONAL NEURAL NETWORK

Fig. 2 shows the architecture of the CNN network [43], [44] used in this study. First, we created an embedding layer, offered by the KERAS framework [45], with a vector space of 8 dimensions and 100 input sequence length. Next, we defined a convolutional layer with parameters 32 filters, ReLu activation function, and a kernel size of 8. In the next layer, we used a pooling layer where the size of the output of the convolutional layer is reduced by half. Then, we flattened the 2D pooled feature maps to one-dimension vector for inserting the output data of the pooling layer into the fully connected sequential dense neural network which is composed of input, hidden, and output layers. The neuron numbers in hidden and output layers are 128 and 6, respectively. We set the dropout rate of the network to 20% and the

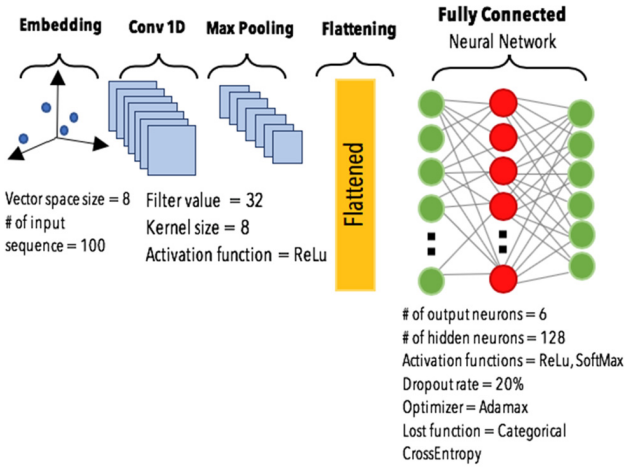

FIGURE 2. CNN architecture developed in this study.

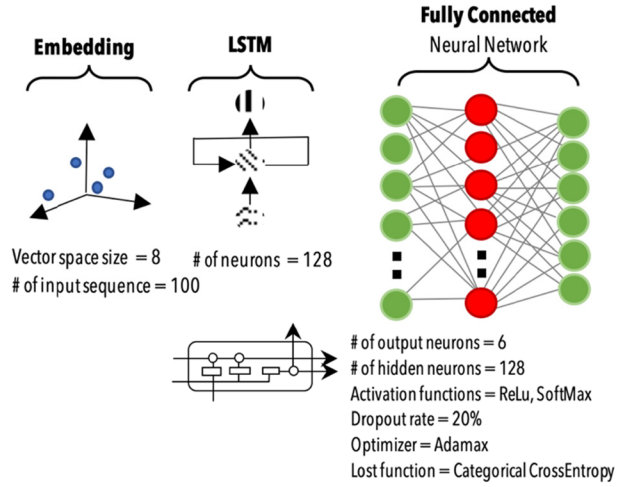

FIGURE 3. RNN architecture developed in this study.

activation functions as ReLu and SoftMax. Besides, we used Adamax and Categorical *CrossEntropy* functions for the lost and optimizer parameters of the network.

3) RECURRENT NEURAL NETWORK AND ITS VARIATION LONG SHORT-TERM MEMORY

RNN is a type of feedforward artificial neural network which can handle variable-length sequence inputs [46]. Unlike traditional feedforward neural networks, RNN uses feedback loops to process sequences in order to maintain memory over time. In the traditional RNN algorithm, recurrent units have very simple structures that have no memory units and additional gates. There is only a simple multiplication of inputs and previous outputs, which is passed through the corresponding activation function. However, an LSTM recurrent unit contains gates, which are used to maintain memory for long periods of time [31].

Fig. 3 represents the architecture of the RNN network used in this study. It is mainly composed of three layers which are embedding, RNN algorithm using LSTM recurrent units, and

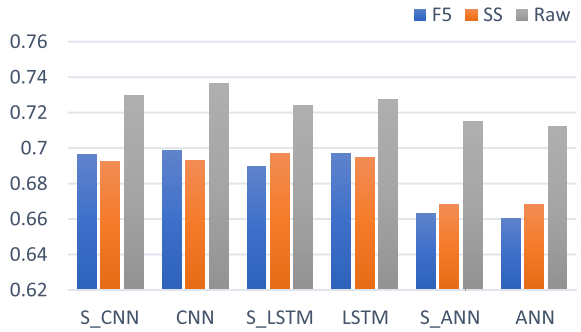

FIGURE 4. Comparison of deep neural network architectures and traditional machine learning methods on TREMO dataset in terms of accuracy.

fully connected sequential dense neural network. We created the embedding layer with a vector space of 8 dimensions and 100 input sequence length. Then for LSTM, we used a number of 128 neurons. At the last layer of our RNN network, we used a fully connected neural network with a hidden layer containing 128 neurons and an output layer composed of 6 output neurons. We set 20% as the dropout rate of the network and the activation functions as ReLu and SoftMax. Besides, we used Adamax and Categorical *CrossEntropy* functions for the lost and optimizer parameters of the network.

IV. EXPERIMENTATION AND RESULTS

To evaluate and compare our proposed architectures, we conducted a set of experiments and analyzed the performance results of three deep learning architectures, ANN, CNN, and LSTM. Besides, we have also examined the effects of different stemming approaches, F5, and SS, in terms of accuracy values [47], [48]. All performance measures are micro-averaged values shown in the results of experiments. In addition, we also investigated the absence and the presence of stop-words. We performed all these experiments on TURTED. However, we have also run the experiments using TREMO dataset for benchmarking and validating proposed architecture. In experimentations, we did not perform cross validation due to computational constraints of large dataset. Thus, we used 90% of the corresponding samples as the training and the rest for testing.

We first examined our proposed deep learning architectures by comparing their performance with traditional machine learning approaches. To do this, we used the TREMO dataset, since it is manually annotated by human judgments. Fig. 4 shows the accuracy results we obtained from the experiments. All proposed architectures resulted in better performances over commonly used machine learning algorithms, support vector machine (SVM), random forest (RF), naïve Bayes (NB), k-Nearest Neighbor (KNN) and decision tree (DT). We have employed hyper-parameter optimization for all machine learning algorithms. For NB, we selected Bayesian optimization using Gaussian process. For SVM, we set the kernel type as linear, cache size as 200 and degree as 3. For RF, number of trees is selected as 10, and

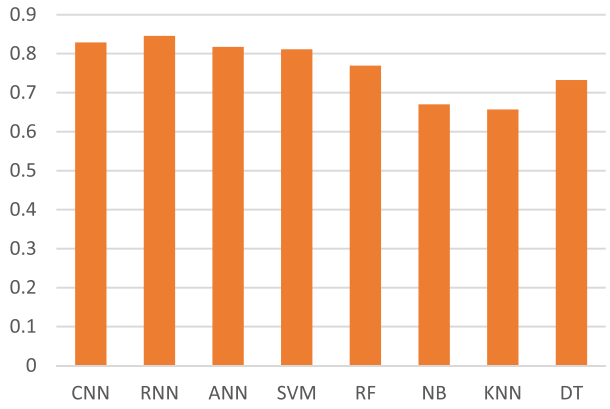

FIGURE 5. Classification results of three algorithms based on different stemming approaches in terms of accuracy values.

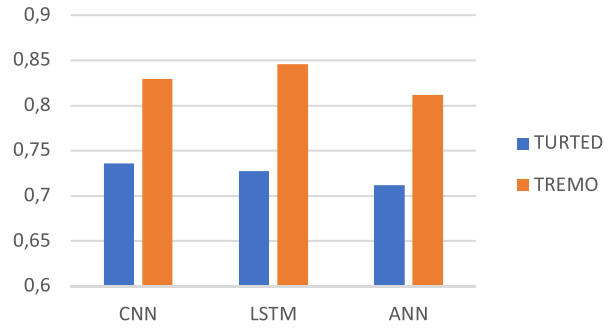

FIGURE 6. Comparison of classification accuracy of three algorithms in TURTED and TREMO.

max depth of a tree is selected. For KNN, we set the number of neighbors as 5 and the distance metric as Minkowski. For DT, the quality split function is set as entropy and maximum depth of the tree is set. Fig. 4 shows that our proposed deep learning architectures are valid and useable for emotion analysis.

Fig. 5 shows a comparison of approaches in terms of accuracy metric, for different deep learning architectures and stemming approaches. Regarding the performance of the algorithms, classification accuracies of CNN outperformed other architectures in most cases. Among the other two architectures, LSTM took the second-highest performance. In Fig. 5, the label names with "S_" prefix indicate the experiments where the stop-words were removed. Generally, the removal of the stop-words from the dataset decreased the results for CNN and LSTM. On the other hand, this is not a case for ANN. This is because CNN and LSTM use the embedding layer, which does not require stop-word elimination. Among the stemming approaches, F5 and SS provided similar results, but the raw dataset, without applying any stemming approach, produced the highest results. This is potentially because of noisy characteristics of Twitter data, where stemming algorithms does not work well.

Fig. 6 presents the comparison of accuracy results of three approaches on two datasets: TURTED and TREMO. CNN and LSTM architectures performed slightly higher results than the ANN algorithm for both datasets. This is because

TABLE 4. Confusion matrix result of the raw TURTED based on six emotion categories.

	Fear	Happy	Disgust	Anger	Sadness	Surprise	Accuracy
Fear	**2,355**	159	132	242	189	176	72.40
Happy	181	**2,026**	134	152	76	156	74.35
Disgust	166	156	**2,354**	209	196	75	74.59
Anger	306	198	207	**3,264**	267	160	74.15
Sadness	243	124	196	302	**2,385**	108	71.02
Surprise	173	120	83	110	79	**1,804**	76.15

TABLE 5. LSTM confusion matrix result of the raw turted based on six emotion categories.

	Fear	Happy	Disgust	Anger	Sadness	Surprise	Accuracy
Fear	**2,367**	173	151	331	221	131	70.15
Happy	168	**1,863**	190	229	98	144	69.21
Disgust	137	139	**2,253**	265	261	48	72.61
Anger	208	173	164	**3,378**	269	96	78.78
Sadness	207	104	171	402	**2,481**	97	71.66
Surprise	149	171	61	196	97	**1,673**	71.28

CNN and LSTM architectures have more layers between input and output layers to construct richer intermediate representations. In general, experiments with the TREMO dataset substantially outperformed the experiments with TURTED. This is because TREMO was manually annotated corpus based on six emotion categories. Furthermore, it syntactically checks for grammatical errors. On the other hand, in the construction of TURTED, the corpus was automatically collected from Twitter by using TEL. There had been no manual validation process so that it inherently includes very noisy data including spelling errors, typos and improper usage of acronyms.

Table 4, Table 5 and Table 6 present confusion matrices showing the distribution of the documents among six emotion categories for the result obtained by using CNN, LSTM, and ANN algorithms, respectively. In each matrix, rows represent the annotations and columns represent the predicted values as returned by the classifier. Considering the results in the three tables, the accuracy performance of the *fear* emotion category is the poorest among the others. On the other hand, the highest performances differentiate according to emotion categories. For example, in Table 4, the *disgust* emotion category performed the highest accuracy result with a value of 74.59%, *anger* emotion category outperformed others in Table 5. In these three confusion matrices, we can also observe the confusion of emotion categories to each other.

Table 4, *fear* and *anger* emotion categories are the most confused categories. For example, 242 documents with *fear* annotation classified into *anger* category. This is expected because TURTED *fear* category includes 33,144 documents in total, where 30453 of them includes TEL keywords belong to *fear* category only. Rest of the 2691 documents in the *fear* category includes keywords belonging to more than one category, and 727 of them includes keywords from *anger* category, in where 656 of them has the second highest MI value

TABLE 6. ANN confusion matrix result of the raw TURTED based on six emotion categories.

	Fear	Happy	Disgust	Anger	Sadness	Surprise	Accuracy
Fear	**2,252**	130	281	273	177	162	68.76
Happy	184	**1,825**	179	231	79	143	69.10
Disgust	163	145	**2,312**	283	157	84	73.54
Anger	239	179	343	**3,183**	297	165	72.24
Sadness	238	91	290	359	**2,296**	146	67.13
Surprise	152	97	69	131	80	**1,853**	77.79

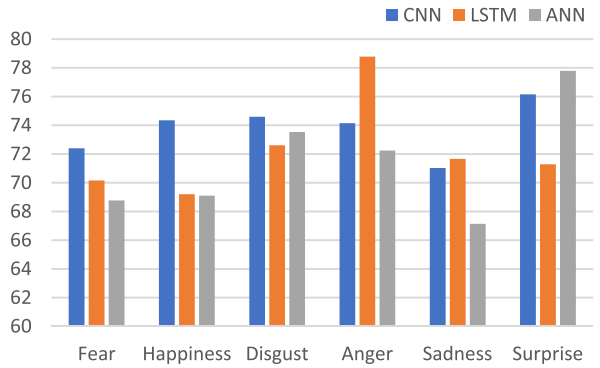

FIGURE 7. Comparison of average accuracy values of each emotion category for CNN, LSTM, and ANN algorithms on TURTED.

for *anger* category. This distribution is also similar for other categories as well. Accordingly, the CNN model produced similar confusion values to the annotation of TURTED.

Fig. 7 depicts the comparison of accuracy measure of each emotion category for three architectures on TURTED individually. There is no one architecture which provides the highest accuracy value for all emotion category. For instance, the ANN algorithm performed the highest accuracy value for the *surprise* emotion category, whereas the performance of CNN is the highest for the *sadness* emotion category. Regarding the performance of each emotion category, the *fear* category performed the lowest accuracy values for all three algorithms at the same time.

V. CONCLUSION

With the rapid increase in social media usage, it became very important to analyze data in terms of emotion categories. Hence, automatic recognition of emotions in text documents is a challenging issue, so labeled datasets play a crucial role in the use of machine learning approaches for this issue. In this paper, we proposed a lexicon-based approach for automatic annotation of texts and curated a dataset, which we named as the Turkish Twitter emotion dataset. The dataset includes more than 195K documents in six emotion categories of *fear*, *happiness*, *disgust*, *anger*, *sadness*, and *surprise*.

After data curation, we developed three deep learning architectures: ANN, CNN, and LSTM. First, we compared the developed architectures on the TREMO dataset with traditional machine learning methods. We showed that the proposed architectures were effective for emotion classification. Then we investigated their performance on TURTED, where we achieved the highest classification performance by using CNN architecture, which was 0.74 in terms of accuracy. Furthermore, we also examined the classification performance of the developed architectures on TURTED and TREMO comparatively. The results showed that experiments in TREMO substantially performed better than the ones with TURTED. This might be the reason of TREMO includes manually corrected text for grammatical errors, contrary to TURTED which includes short text with a large number of potential spelling errors and typos.

In this study, all proposed deep learning architectures outperformed traditional machine learning algorithms for emotion classification in Turkish. In general terms, deep neural networks scale better with more data than traditional machine learning algorithms. It is already known that the use of more data improves the accuracy with deep neural networks. Besides, another important reason is the usage of embedding layer to extract valuable features in deep learning architectures. On the other hand, traditional machine learning algorithms often require complex feature engineering, where stemming can be considered as a part of it. However, our experimentation showed that stemming did not work well with Twitter data, which is very noisy and includes very strange informal language with too much spelling errors. This might explain why traditional machine learning approaches produced lower results. Conversely, deep neural network does not need stemming since it inherently learns stemming in word-embedding layer on actual data.

For future work, it might be worth to extend the study in several dimensions: First, the size of the automatically constructed dataset can be enlarged. Second, several word embedding schemes such as word2vec, fastText, gloVe, and LDA2Vec can be used and their performances might be compared.

REFERENCES

[1] P. Ekman, "Are there basic emotions?" *Psychol. Rev.*, vol. 99, no. 3, pp. 550–553, Jul. 1992.
[2] W. James, "What is an emotion?" *Mind*, vol. 9, no. 34, pp. 188–205, Apr. 1884.
[3] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.
[4] S. M. Mohammad, "#emotional tweets," in *Proc. 1st Joint Conf. Lexical Comput. Semantics (SEM)*, Montréal, QC, Canada, 2012, pp. 246–255.
[5] I. Ileri and P. Karagoz, "Detecting user emotions in Twitter through collective classification," in *Proc. Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manage. (IC3K)*, Porto, Portugal, 2016, pp. 205–212.
[6] S. Demirci, "Emotion analysis on Turkish tweets," M.S. thesis, Dept. Comp. Eng., Middle East Tech. Univ., Ankara, Turkey, 2014.
[7] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," in *Proc. 6th Joint Conf. Lexical Comput. Semantics (SEM)*, Vancouver, BC, Canada, vol. 2017, pp. 65–77.
[8] W. A. Hussien, Y. M. Tashtoush, M. Al-Ayyoub, and M. N. Al-Kabi, "Are emoticons good enough to train emotion classifiers of Arabic tweets?" in *Proc. 7th Int. Conf. Comput. Sci. Inf. Technol. (CSIT)*, Amman, Jordan, 2016, pp. 1–6.
[9] M. Hasan, E. Rundensteiner, and E. Agu, "EMOTEX: Detecting emotions in Twitter messages," in *Proc. ASE Bigdata/Socialcom/Cybersecurity Conf.*, Stanford, CA, USA, May 2014.
[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proc. Workshop Lang. Social Media*, Portland, OR, USA, 2011, pp. 30–38.

[11] R. Velioglu, T. Yildiz, and S. Yildirim, "Sentiment analysis using learning approaches over emojis for Turkish tweets," in *Proc. 3rd Int. Conf. Comput. Sci. Eng. (UBMK)*, Sarajevo, Bosnia, 2018, pp. 303–307.
[12] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the OMG!" in *Proc. 5th Int. Conf. Weblogs Social Media*, Catalonia, Spain, 2011, pp. 538–541.
[13] M. A. Tocoglu and A. Alpkocak, "Emotion extraction from turkish text," in *Proc. Eur. Netw. Intell. Conf. (ENIC)*, Wrocław, Poland, 2014, pp. 130–133.
[14] G. Gautam and D. Yadav, "Sentiment analysis of Twitter data using machine learning approaches and semantic analysis," in *Proc. 7th Int. Conf. Contemp. Comput. (IC3)*, Noida, India, 2014, pp. 437–442.
[15] S. A. Bahrainian and A. Dengel, "Sentiment analysis using sentiment features," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT)*, Atlanta, GA, USA, Nov. 2013, pp. 26–29.
[16] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA, vol. 1, 2011, pp. 151–160.
[17] E. M. van den Broek-Altenburg and A. J. Atherly, "Using social media to identify consumers' sentiments towards attributes of health insurance during enrollment season," *Appl. Sci.*, vol. 9, p. 10, May 2019.
[18] M. Jabreel and A. Moreno, "A deep learning-based approach for multi-label emotion classification in tweets," *Appl. Sci.*, vol. 9, p. 6, Mar. 2019.
[19] Z. Boynukalın, "Emotion analysis of Turkish texts by using machine learning methods," M.S. thesis, Dept. Comp. Eng., Middle East Tech. Univ., Ankara, Turkey, 2012.
[20] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern Recognit. Lett.*, vol. 93, pp. 133–142, Jul. 2017.
[21] Ö. Çoban, B. Özyer, and G. T. Özyer, "Sentiment analysis for Turkish Twitter feeds," in *Proc. 23nd Signal Process. Commun. Appl. Conf. (SIU)*, Malatya, Turkey, 2015, pp. 2388–2391.
[22] E. S. Akgül, C. Ertano, and B. Diri, "Sentiment analysis with Twitter," *Pamukkale Univ. J. Eng. Sci.*, vol. 22, no. 2, pp. 106–110, 2016.
[23] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *J. Personality Social Psychol.*, vol. 66, no. 2, pp. 310–328, 1994.
[24] M. A. Tocoglu and A. Alpkocak, "TREMO: A dataset for emotion analysis in Turkish," *J. Inf. Sci.*, vol. 44, no. 6, pp. 848–860, Dec. 2018.
[25] I. Ileri, "Collective classification of user emotions in Twitter," M.S. thesis, Dept. Comp. Eng., Middle East Tech. Univ., Ankara, Turkey, 2015.
[26] R. Plutchick, "Emotions: A general psychoevolutionary theory," in *Approaches to Emotion*, K. R. Scherer and P. Ekman, Eds. New York, NY, USA: Taylor & Francis, 2009.
[27] S. M. Mohammad and S. Kiritchenko, "Understanding emotions: A dataset of tweets to study interactions between affect categories," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, Miyazaki, Japan, 2018.
[28] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text," in *Proc. 4th Int. Workshop Semantic Eval. (SemEval)*, Prague, Czech Republic, 2007, pp. 70–74.
[29] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing Twitter 'big data' for automatic emotion identification," in *Proc. Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Social Comput.*, Amsterdam, The Netherlands, 2012, pp. 587–592.
[30] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Proc. Int. Conf. Text, Speech Dialogue (TSD)*. Berlin, Germany: Springer, 2007, pp. 196–205.
[31] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *J. Comput. Sci.*, vol. 36, Sep. 2019, Art. no. 101013, doi: 10.1016/j.jocs.2019.05.009.
[32] A. Qadir and E. Riloff, "Bootstrapped learning of emotion hashtags #hashtags4you," in *Proc. 4th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Atlanta, GA, USA, 2013, pp. 2–11.
[33] M. Hasan, E. Rundensteiner, and E. Agu, "Automatic emotion detection in text streams by analyzing Twitter data," *Int. J. Data Sci. Anal.*, vol. 7, no. 1, pp. 35–51, 2019.
[34] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, and O. M. Vursavas, "Information retrieval on turkish texts," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 3, pp. 407–421, Feb. 2008.
[35] K. Oflazer, "Two-level description of turkish morphology," *Literary Linguistic Comput.*, vol. 9, no. 2, pp. 137–148, Jan. 1994.
[36] M. A. Toçoglu and A. Alpkoçak, "Lexicon-based emotion analysis in Turkish," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 27, no. 2, pp. 1213–1227, Apr. 2019.
[37] J. Roesslein. *Tweepy Documentation, Revision ce2ad288*. Accessed: 2019. [Online]. Available: http://docs.tweepy.org/en/latest/
[38] M. Civriz, "Dictionary-based effective and efficient Turkish lemmatizer," M.S. thesis, Dept. Comp. Eng., Dokuz Eylül Univ., Izmir, Turkey, 2011.
[39] M. F. Porter. (Oct. 2001). *Snowball: A Language for Stemming Algorithms*. [Online]. Available: http://snowball.tartarus.org/texts/introduction.html
[40] Y. Bengio, "Learning deep architectures for AI," *J. Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–27, Jan. 2009.
[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[42] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for Twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23253–23260, Jan. 2018.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, vol. 1, 2012, pp. 1097–1105.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[45] S. Pal and A. Gulli, *Deep Learning with Keras*. Birmingham, U.K.: Packt, 2017.
[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[47] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, MA, USA: MIT Press, 2016.
[48] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

MANSUR ALP TOCOGLU received the B.Sc. degree in software engineering and the M.Sc. degree in artificial intelligent systems from the Izmir University of Economics, İzmir, Turkey, in 2008 and 2013, respectively, and the Ph.D. degree in computer engineering from Dokuz Eylül University, İzmir. He has been working as a Research Assistant with the Software Engineering Department, Manisa Celal Bayar University, Manisa, Turkey. His research interest includes information extraction from text using machine learning techniques.

OKAN OZTURKMENOGLU received the B.Sc. degree in computer engineering from Selçuk University, Konya, Turkey in 2008, and the M.Sc. and Ph.D. degrees in computer engineering from Dokuz Eylül University, İzmir, Turkey, in 2012 and 2018, respectively. He is working as a Research Assistant with Dokuz Eylül University. His research areas focus about information retrieval and information extraction, such as named entity recognition and question answering systems.

ADIL ALPKOCAK received the B.Sc. degree from Hacettepe University, Ankara, and the M.Sc. and Ph.D. degrees in computer engineering from Ege University, in 1992 and 1998, respectively. He has been currently working as an Associate Professor with the Department of Computer Engineering, Dokuz Eylül University. He is also the Founder and a Coordinator of the Dokuz Eylul Multimedia Information Retrieval (DEMIR) Research Laboratory. His research interests include multimedia information retrieval, text mining, information extraction, and their implementations of research topics in medical informatics.