Article

JIS

Journal of Information Science
2018, Vol. 44(6) 848–860

DOI: 10.1177/0165551518761014
journals.sagepub.com/home/jis
SAGE

TREMO: A dataset for emotion analysis in Turkish

Mansur Alp Tocoglu
Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, Turkey

Adil Alpkocak
Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, Turkey

Abstract
This study presents a new dataset to be used in emotion extraction studies in Turkish text. We consider emotion extraction as a supervised text classification problem, which thereby requires a dataset for the training process. To satisfy this requirement, we aim to create a new dataset containing data for the six emotion categories: happiness, fear, anger, sadness, disgust and surprise. To gather this dataset, we conducted a survey and collected 27,350 entries from 4709 individuals. In the next step, we performed a validation process in which annotators validated each entry one by one by assigning a related emotion category. As a result of this process, we obtained two datasets, one raw and the other validated. Subsequently, we generated four versions of these two datasets using two different stemming methods and then modelled them using a vector space model. Then, we ran machine learning algorithms, including complement naive Bayes (CNB), random forest (RF), decision tree C4.5 (J48) and an updated version of support vector machines (SVMs), on the models to calculate the accuracy, precision, recall and *F*-measure values. Based on the results we obtained, we concluded that the SVM classifier yielded the highest performance value and that the models trained with a validated dataset provide more accurate results than the models trained with a non-validated dataset.

Keywords
Emotion analysis; emotion extraction; text classification; TREMO dataset; Turkish language

1. Introduction

The increasing use of social media applications has led to the extraction of a huge amount of raw text data. This raises the problem of extracting meaningful data from large amounts of raw data. Extracting meaningful information from such non-structured data requires applying very complex and expensive processes. Consequently, many kinds of classification algorithms have been developed in the literature to overcome this problem. The primary purpose of these algorithms is to categorise textual data with similar structures and meanings to create different categories. These newly created category groups can be used to categorise new non-structural text files.

Social media tools such as Twitter and Facebook play an important role as sources of big data in the process of extracting information from any text. The most important reason for this is that the text data being generated in these applications is increasing day by day in large quantities. However, since these resources are not categorised, they cannot meet the requirements of classification algorithms. In the literature, there are many datasets created to solve this problem. One of them is the TTC-3600 benchmark dataset formed for Turkish text categorisation [1]. In addition, studying emotion extraction problem in Turkish texts requires a categorised dataset in this area, and creating such a dataset is the goal of this study.

In this article, we present a new dataset called TREMO, which is publicly available for the use of academic researchers.[1] To the best of our knowledge, TREMO is the first generated dataset for emotion analysis processes in Turkish. It is

Corresponding author:
Mansur Alp Tocoglu, Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, 35390 Izmir, Turkey.
Email: mansur.tocoglu@ceng.deu.edu.tr

based on a survey with 5000 participants asked to write stories about their strongest experienced memories for the basic six emotions. After eliminating unacceptable replies, we accepted 4709 individuals and obtained 27,350 entries in total. In the next step, the dataset is passed on to annotators who classified each entry with an emotion category. In this process, an entry is considered as validated when three annotators agree on the same emotion category, where some of the entries not complying with this condition were removed from the dataset. Then, we applied a set of machine learning algorithms to evaluate the effects of the validation process and feature selection methods.

The remainder of the article is organised as follows. In section 'Related work', we discuss the datasets in the literature, which are collected for emotion analysis. In section 'TREMO dataset preparation', we describe how the raw data is collected, and we share basic statistical information about it. In section 'Validation of the dataset', we share the steps taken within the validation process of the dataset. In section 'Experiments on datasets', we share detailed information regarding the steps taken before running the classification algorithms on the datasets and the results obtained using these classification algorithms. In section 'Conclusion and future work', we conclude the article and mention about the future studies.

2. Related work

In the literature, there are several datasets created for emotion classification in English. One of the best known among these corpora is the English-oriented International Survey on Emotion Antecedents and Reactions (ISEAR) dataset [2]. In total, 3000 volunteers from 37 different countries participated in this project by writing about their life experiences and reactions for seven major emotions: joy, fear, anger, sadness, disgust, shame and guilt. This dataset is used in some studies for emotion classification [3,4]. Giachanou and Crestani [5] conducted a study about Twitter opinion retrieval in which they retrieved tweets related to a particular topic. They used several datasets to achieve this goal, one of which was the dataset created by Luo et al. [6], which is composed of 50 topics and 5000 tweets. Their second dataset was the AFINN Lexicon, which consists of more than 2000 words to be used in identifying opinion-related terms. Go et al. [7] worked on classifying tweets as either negative or positive. They created their dataset by gathering tweets according to emoticons that indicate the differences between positive and negative emotions. Mohammad [8] focused on determining whether word-emotion association lexicons yield better results than using *n*-gram features. In addition, he found that emotion lexicon features yield better results in new domains than using *n*-gram features. To achieve these steps, he used two-emotion lexicon features annotated for Ekman's six emotions, WordNet Affect Lexicon [9] and NRC-10 [10,11]. For his training dataset, he chose to use the SemEval-2007 Affective Text corpus [12]. Chaffer and Inkpen [13] extracted six emotions from a text using a dataset that is heterogeneous by subjects such as news headlines, fairy tales and blogs. They adopted supervised machine learning techniques and features, such as bag-of-words and N-grams. Kouloumpis et al. [14] evaluated effects of using features on Twitter sentiment analysis using supervised learning approaches. To achieve this goal, they used three different corpora of Twitter messages. Two of these, hashtagged and emoticon datasets, were used as training datasets. For testing their model, they benefitted from a dataset that was annotated manually. Yang et al. [15] worked on emotion classification problems for four emotion categories: joy, happiness, sadness and fear. They used blog posts and emoticons as training datasets and focused on comparing the results obtained by support vector machines (SVMs) and conditional random field classifiers.

There is no dataset publicly available for emotion analysis in Turkish. Most of the studies used non-Turkish datasets translated into Turkish, such as Boynukalin used a portion of the ISEAR dataset containing documents for four emotions translated into Turkish by 33 people in her study of extracting emotions from Turkish text. In addition to this translated corpus, Boynukalin [16] also used 25 children's Turkish fairy tales as dataset in her study. In other studies, datasets are collected from social media tools. Demirci studied emotion extraction from Turkish micro-blog entries. She focused on gathering tweets for the six emotions: anger, disgust, fear, joy, sadness and surprise, using the Twitter search mechanism for hashtags. For each emotion category, Demirci [17] defined hashtags containing the derivatives of each emotion word. As a result, Demirci succeeded in collecting 1000 tweets for each emotion, 6000 tweets in total.

To the best of our knowledge, TREMO is the first dataset created in Turkish for emotion analysis including six emotion categories: happiness, fear, anger, sadness, disgust and surprise. In addition, it is also the first dataset annotated, which is validated through external annotators.

3. TREMO dataset preparation

We conducted a survey with the participation of 5000 people from different living areas and different age ranges to collect a dataset based on six emotions. In this survey, we asked participants to share their memories or any experiences they would have for the six emotion categories that Ekman [18] described. As a result of this process, a total of 4709 people were approved for participation, and a total of 27,350 entries were collected. Participation in this survey was conducted

Table 1. Distribution of attendance types in the survey.

Attendance type	Attendance number	Female participants	Male participants
Web-based	673	392	281
Paper-based	4036	2378	1658
Total	4709	2770	1939

Table 2. Distribution of individuals in the survey according to age groups.

Age ranges (years)	No. of individuals
14–20	3854
21–30	531
31–40	157
41–50	101
51–70	60

either through a website or by manually filling in the fields for each emotion category in a given paper. In Table 1, the participation rate using the first method is very low compared with that of the second method. The most important reason for this is that we were not able to obtain the number of participants we had planned to collect on the web-based method. So, we also conducted the survey at high schools and universities. Of course, in this case, we had to give each participant a paper to fill out. This put an additional burden on us, as we had to enter each paper into the system. The same table also shows the female–male distributions of the participants. There were more female participants than males.

The ages of most of the participants were between 15 and 24 years as a result of having many participants from high schools and universities. Table 2 shows the distribution of the participants in five different age groups. Most participants were positioned in two age groups, 14–20 and 21–30, that comprise the normal age range for a student.

As noted, we conducted this survey at several universities and high schools. The majority of these are educational institutions located in Izmir. There are also high schools in different cities in Turkey, including Ankara, Balikesir and Diyarbakir, where we conducted the survey. We obtained official permission from the Provincial Directorate of National Education to be able to go to the high schools in Izmir. While making this choice, we focused on visiting high-ranked schools. In addition to high schools, we also went to two state universities, Dokuz Eylul University and Izmir Katip Celebi University. Because we went to many educational institutions, 73.88% of the participants were high school students and 15.44% were university students. The remaining participants practised 32 different occupations.

In the survey, some participants could not manage to write an entry for each emotion category. This caused differences in the distribution of the entries for each emotion, which is shown in Table 3.

4. Validation of the dataset

The validation process of the dataset plays an important role in this study. The most important reason for this is the elimination of entries that are considered ambiguous or fake in terms of emotional categorisation. If such entries are not discarded from the raw dataset, they can result in many outliers being in the training set. This negatively impacts the results obtained from supervised learning algorithms. In the validation process, we first created an application supporting web and mobile interfaces. Annotators who want to join the validation process first register on the system, giving basic information such as name, surname, gender, occupation, age, email address and define a password which is required by the system during annotation process. In the following stage, the annotator enters the system only after the system administrator's approval. This is designed to prevent unauthorised registrations on the system to secure the validation process. An annotator who obtains confirmation can log in to the system using his or her specified mail address and password. At annotation stage, each entry is displayed in random order to annotator. The annotator simply clicks a button, representing one of the six emotion categories, to annotate the entry. In addition, an extra button is placed for ambiguous condition, which is used when the annotator cannot decide on suitable emotion category.

Each entry is presented to at least three different annotators. If three of them annotate the same emotion category, then we assume that the entry is validated. If not, then the entry is presented to different annotators until reaching three

Table 3. Distribution of the entries for each emotion category.

Happiness	Fear	Anger	Sadness	Disgust	Surprise
4700	4616	4636	4664	4522	4212

Table 4. Distribution of the validation conditions of the entries at the end of the validation process.

Conditions	Votes	No. of validated entries	No. of entries validated with their original emotion	No. of entries validated with different emotion
Consensus	3-0	19,462	18,154	1308
Majority-of-votes	3-1	4583	3639	944
	3-1-1, 3-2	1944	1190	754
Reject	–	1361	0	1361

Table 5. Examples of entries translated into English with their original and validated emotion categories, conditions and vote distributions in the validation process.

ID	Entry (English Translation)	Original emotion	Validated emotion	Condition	Vote distribution
6	I am surprised to encounter a surprise that I never expected	Surprise	Surprise	Consensus (3-0)	3 Surprise
2048	My colleague's attitude is bothering me	Disgust	Anger	Consensus (3-0)	3 Anger
3254	Little gestures from a stranger and behaviours that prove that he or she cares about you.	Happiness	Happiness	Majority-of-votes (3-1)	3 Happiness 1 Surprise
116	The moment I notice that I have uploaded the wrong assignment	Fear	Fear	Majority-of-votes (3-1-1)	3 Fear 1 Sadness 1 Anger
3741	In general, I sleep by opening the television and cutting down the volume of it	Fear	Ambiguous	Reject	2 Happiness 2 Fear 1 Ambiguous
19301	When I see a goalkeeper scored to his own goal	Surprise	Ambiguous	Reject	1 Happiness 1 Sadness 1 Surprise 1 Ambiguous 1 Anger

votes for the same emotion category. If three votes are not obtained at the end of five annotations, we remove corresponding entry from the dataset. Figure 1 shows basic steps in validation of an entry. In the validation process, there are three different possible conditions, which are consensus, majority-of-votes and reject. In this process, if consensus is reached with first three annotators, system makes a decision. Majority-of-votes may have three different vote distributions (i.e. 3-1-1, 3-1 and 3-2 votes), which is also reaching three votes for the same emotion category. On the other side, reject condition has also three different vote distributions (i.e. 2-2-1 or 1-1-1-1-1 and 2-1-1-1 votes). In addition, it is also possible to reject entries with consensus and majority-of-votes conditions since an annotator can make a choice for ambiguous condition. Table 4 shows the number of entries in all these conditions after the validation process. Table 5 shows the examples of entries translated into English with their original and validated emotion categories, conditions and vote distributions in the validation process.

In validation process, 48 volunteered annotators worked for 92,986 individual annotations. Annotators have 11 different professions where engineers, students and academicians are among the prominent, and their age distribution is between 14 and 67 years, where only two of them are less than 20 years old. The maximum number of individual annotations, which an annotator performed, is 10,763, and the minimum is only 8.

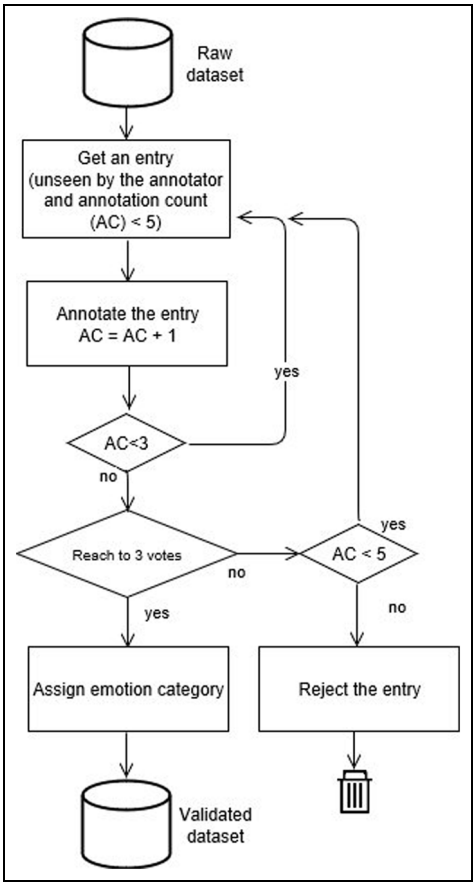

Figure 1. Basic steps in the validation of an entry (initial value of AC is 0).

Table 6. Distribution of the individual annotations according to annotators' age ranges.

Age ranges (years)	No. of annotations
14–20	790
21–30	37,103
31–40	20,446
41–50	1317
51–60	19,119
61–70	14,211

Table 6 shows the total numbers of individual annotations versus age groups of annotators in the validation process. Two of these age groups, 21–30 and 31–40 years, are the top two dominant groups since they have the maximum number of annotators. Figure 2 represents the distribution of annotators' contribution to validation process, which illustrates the proportion of the total annotations that is cumulatively annotated by percentage of the annotators. For example, the top 25% of annotators has 79% of the whole individual annotations in the validation process. Furthermore, we evaluated level of agreement between annotators by Cohen's [19] kappa value which is found 0.83 indicating very good level of agreement between annotators.

The validation process discards entries containing ambiguity in their emotion categories. As a result, we removed 1361 entries in total, comprising 4.98% of the overall raw dataset, and called this new version as validated dataset. Table 7 represents the total number of entries both original and the resulted number after the validation process, where a clear difference can be easily observed. Some of the entries were annotated contradictorily to participant's original emotional category. This indicates that the raw dataset includes some fake entries, or some of the emotions such as surprise and

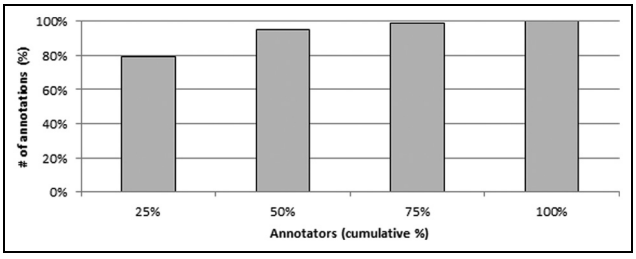

Figure 2. Cumulative percentage of annotations versus cumulative share of annotators.

happiness are easy to confuse. In addition, a decrease in the number of surprise entries is clearly observable while happiness is increased.

Table 8 represents a table of confusion in raw versus validated datasets based on emotional categories. It shows how original emotions are interpreted and annotated differently in the validation process. For example, 642 entries originally categorised as surprise in the raw dataset are annotated as happy.

5. Experiments on datasets

In this section, we describe classification experiments that we performed on TREMO, including both raw and validated datasets. The difference between these two datasets is the elimination of the 1361 ambiguous entries defined in the validation process. Before the experiments are conducted, we pre-processed datasets to remove unnecessary structures and prepared them to be used as training datasets. In the first step of the pre-processing, four dataset versions are generated from the two relevant datasets (raw and validated) using two stemming methods. Next, ineffective terms and numerical values are deleted. After the completion of pre-processing, feature selection is performed to eliminate insignificant features and minimise the dimensions of the dataset versions. Then, we transformed these versions to vector space models, where document term matrices (DTMs) are generated. In section 'Experimental results', we present classification results we obtained using four machine learning techniques.

5.1. Pre-processing

The goal of pre-processing is to prepare the dataset to classification process. The first step is to find the stems of the terms in each entry, then, we deleted the punctuation marks, the extra spaces, the numeric characters and the fluff terms from these datasets. For this purpose, we used two different stemming methods, fixed prefix stemming (FPS) [20] and a directory-based Turkish stemmer named Zemberek (Z) [21]. FPS simply gets the first n characters of a term, trims out the rest. At this point in the study, we chose n as five to represent the first five characters (F5). We chose F5, instead of F4 or F7, since it has been shown that it has optimum performance in terms of effectiveness measures among others [20]. We performed these two stemming methods upon the raw and validated datasets and created four different dataset versions. Table 9 shows some numeric properties of all four versions of TREMO, where V character in the name of dataset versions represents validated dataset while the others are for raw dataset.

5.2. Feature selection

After the creation of four dataset versions based on two stemming methods, we used feature selection to remove insignificant features (terms) and to reduce the dimensionality of the feature sets [22]. In this study, applying a feature selection method to the datasets plays an important role because of the high number of features within the datasets that prevent running some classification algorithms. For ranking the most significant terms for each emotion category, we decided to use mutual information (MI) [23]. We reordered the features of these four dataset versions from the highest to the lowest MI value and then selected the most valuable features for each emotion category based on two approaches. The first is the selection of the first 500 features, and the other one is to make the selection for a given threshold. Here, we empirically chose a threshold value of MI value of the 500th feature of the happiness emotion category. This is because we obtained one of the highest classification results using complement naive Bayes (CNB) [24] as the classifier and F5 as the training dataset. Table 10 shows the threshold values for the four dataset versions and the number of features fit into the given

Table 7. Distribution of the entries after the validation process.

Emotion category	Original no. of entries	No. of entries after the validation process
Happiness	4700	5229
Fear	4616	4393
Anger	4636	4723
Sadness	4664	5021
Disgust	4522	3620
Surprise	4212	3003
Total	27,350	25,989

Table 8. Table of confusion in raw versus validated datasets based on emotional categories.

		Validated dataset							
		Happy	Fear	Anger	Sadness	Disgust	Surprise	Reject	Total
Raw dataset	Happy	4513	15	2	14	1	59	96	4700
	Fear	19	4049	51	246	21	26	204	4616
	Anger	19	35	3934	357	24	35	232	4636
	Sadness	20	95	186	4101	11	33	218	4664
	Disgust	16	151	421	48	3552	16	318	4522
	Surprise	642	48	129	255	11	2834	293	4212
	Total	5229	4393	4723	5021	3620	3003	1361	27,350

Table 9. Numerical properties of the four dataset versions of TREMO.

Dataset	Dataset versions	Total entry	Total terms	Unique terms
Raw	F5	27,350	132,485	6489
	Z	27,348	129,267	4142
Validated	F5_V	25,989	126,593	6280
	Z_V	25,989	123,581	4009

Table 10. Number of selected features for each emotion category and threshold values for each version.

Emotion types/threshold	F5	F5_V	Z	Z_V
Happiness	500	517	535	500
Fear	509	498	504	450
Anger	522	527	484	477
Sadness	399	409	408	397
Disgust	646	562	618	547
Surprise	371	292	409	310
Threshold value	0.000234	0.0002663	0.000185	0.00021

threshold value for each emotion category. To avoid the repetition of features within each emotion category, we took intersection of the features. The intersected feature numbers according to each dataset version are shown in Table 11.

Table 12 presents numeric information of the 10 highest MI-valued terms of the Z dataset version for the happiness category. The term column includes English translations in parenthesis. In addition to MI value, it also shows the overall frequency values, the ratio of the frequency value to the total number of entries and the value indicating the number of entries having the related term in their contents.

Table 11. Number of intersected features for each feature selection approaches for four dataset versions.

Dataset versions	Feature selection approaches	
	First 500 terms	Threshold
F5	1439	1395
F5_V	1397	1336
Z	1336	1386
Z_V	1316	1192

Table 12. Numerical information of the 10 highest MI-valued terms of the Z dataset version for the happiness emotion category, where English translation is given in parenthesis.

Term	MI	Frequency	Frequency/*N*	No. of entries containing the term
mutlu (happy)	0.1965	2095	0.0766	2017
ol (happen)	0.0584	5232	0.1913	4702
kork (fear)	0.0215	2135	0.0781	2102
sevin (glad)	0.0177	235	0.0086	233
şaşır (surprise)	0.0173	1753	0.0641	1740
kazan (win)	0.0169	543	0.0199	534
üzül (sad)	0.0162	1695	0.0620	1641
tiksin (disgust)	0.0146	1474	0.0539	1464
öfkelen (anger)	0.0146	1437	0.0525	1427
birlikte (together)	0.0097	183	0.0067	182

MI: mutual information.

N is the total number of entries.

After finishing pre-processing and feature selection phases, the next step is to transform these dataset versions into a vector space model [25]. In this model, each entry is represented as a vector in DTM. In DTM, each row is a vector where columns represent terms. DTM is a sparse matrix with zero values in most of its cells. The cells contain a non-zero value only when the corresponding term occurs within the corresponding entry. Finally, we used TF $\times$ IDF weighting scheme [23] in vector space model.

5.3. Experimental results

We set up a series of experimentations to determine whether the raw dataset yields better classification results after passing through the validation process. For this reason, we subjected four dataset versions, F5, F5_V, Z and Z_V, to evaluate the performance of different classification algorithms, which are CNB, random forest (RF) [26], decision tree C4.5 (J48) [22] and an updated version of SVMs [27]. All the experiments are implemented in WEKA version 3.6.14 [28]. We implemented 10-fold cross-validation in evaluation of the performance of each classifier, where 90% of dataset is used as a training set, and the remainder is used for testing. We compared classification results in terms of accuracy, precision, recall and *F*-measure. Furthermore, we also examined the reflections of these results by emotion categories.

Figure 3 shows the general classification results obtained for each dataset versions according to the three classification algorithms without using any feature selection approach. Here, we did not include RF algorithm because of the high dimensionality of datasets. Despite minor differences, the figure shows that the two stemming techniques yield similar results. However, it appears that the best classification results are obtained when using SVM as the classification algorithm. Besides, the deletion of ambiguous entries detected in the validation process from the raw dataset has positive effects on the overall average accuracy values by 5.7% of F5 stemmed datasets and by 5.6% of Z stemmed datasets. Apart from that, the CNB and SVM techniques yield slightly better results than the J48 algorithm.

Figure 4 shows overall accuracy values of the raw and validated dataset versions subjected to the first 500 terms feature selection. We compared the classification accuracy results for four classification algorithms. All the comparison results are very close to each other, so there is no clear-cut winner. CNB and SVM produced higher scores for F5 and F5_V than Z and Z_V dataset versions. However, J48 and RF have better results on Z and Z_V dataset versions. These

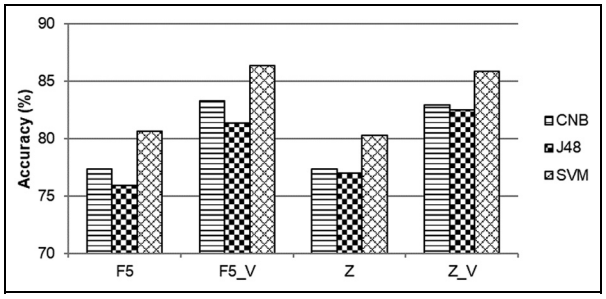

Figure 3. General accuracy results obtained for each dataset version according to three different classifiers.

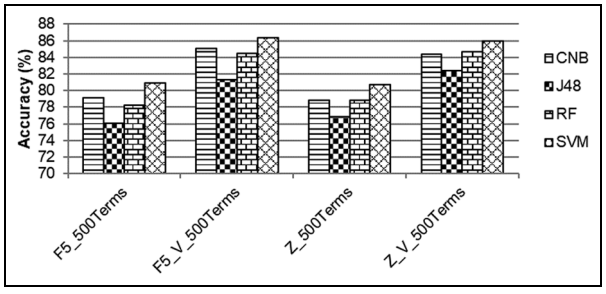

Figure 4. Overall accuracy values of the dataset versions subjected to first 500 terms feature selection approach.

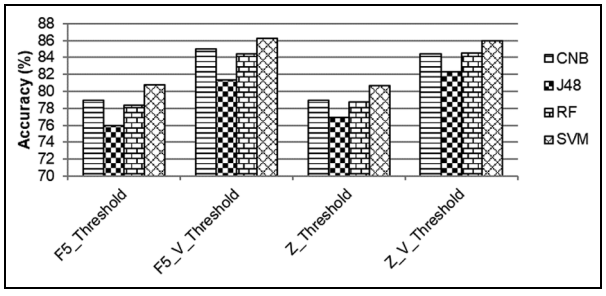

Figure 5. Overall accuracy values of the dataset versions subjected to the threshold feature selection approach.

results explain that both stemming methods have similar results. These results also repeat in threshold feature selection approach, which is shown in Figure 5.

In another experiment, we aim to determine which feature selection approach yields higher classification results for the dataset versions F5 and F5_V. The results we obtained from the experiment are shown in Figure 6. It is obvious that there are no major differences in the results between feature selection approaches. The first 500 terms feature selection approach provides slightly higher results compared with threshold approach for the classifiers CNB, J48 and SVM applied on F5 dataset version. However, when the dataset version F5_V is used, there is a tie between the two feature selection approaches. The first 500 terms feature selection approach provided better results for the classifiers CNB and SVM, and the threshold approach achieved higher results using J48 and RF algorithms.

In Figure 7, overall accuracy values of the three F5_V dataset versions are compared with each other. One of them, F5_V, is not subject to any feature selection approach, and the other two versions are subject to both feature selection approaches. As a result, we concluded that feature selection approaches applied to F5_V dataset version produce slightly better in accuracy values for CNB. However, the accuracy result of J48 is slightly better for the dataset version F5_V that is not subject to any selection approach.

Tables 13 and 14 show the confusion matrix results of the models trained using the F5 and F5_V dataset versions for SVM classifier. In both tables, the rows represent the ground truth data, and the columns represent the classifier results.

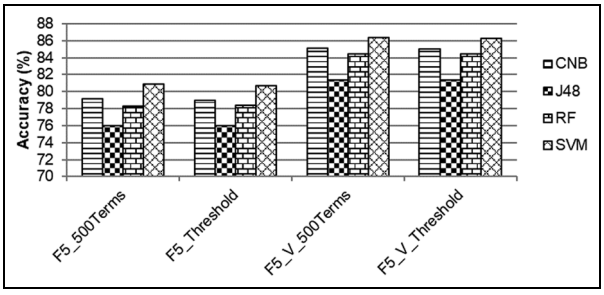

Figure 6. Overall accuracy values of the dataset versions F5 and F5_V subjected to two feature selection approaches.

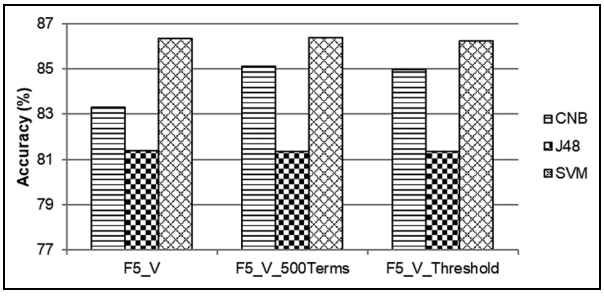

Figure 7. Overall accuracy values of all F5_V dataset versions.

Table 13. Confusion matrix of SVM algorithm on F5 dataset version.

	Happiness	Fear	Anger	Sadness	Disgust	Surprise	Accuracy
Happiness	4052	99	117	100	57	**275**	**86.21**
Fear	164	3758	164	**335**	124	71	81.41
Anger	155	151	3653	**305**	**227**	145	78.80
Sadness	266	186	**328**	3630	78	176	77.83
Disgust	92	132	**311**	90	3832	65	84.74
Surprise	**419**	126	235	222	85	3125	74.19

SVM: support vector machine.

The last column of these tables provides the individual accuracy values of each emotion category. For the raw dataset version F5, the accuracy value of the happiness emotion category is the highest. For the validated dataset version, F5_V, the disgust emotion category receives the highest score. Furthermore, we marked the most confused emotion classification results in boldface. For example, in Table 13, 275 happiness ground truth entries are classified as entries indicating the surprise emotion category.

Tables 15 and 16 show the precision, recall and *F*-measures of each emotion categories. These results are obtained using the F5 and F5_V dataset versions for SVM classifier. We obtained the highest values for the emotions disgust and happiness categories. The reason for this might be that these two categories also have the highest accuracy values, as shown in the confusion matrices in Tables 13 and 14.

6. Conclusion and future work

The major purpose of this study is to prepare a dataset for emotion analysis with six emotions in Turkish. To do this, we first conducted a survey and obtained 27,350 entries from 4709 individuals. To validate this raw dataset, we performed a validation process where 48 annotators voluntarily participated. Here, a total of 92,986 individual annotations were made, and at the end, 1361 entries were discarded from the raw dataset owing to ambiguities in the emotion categories. After

Table 14. Confusion matrix of SVM algorithm on F5_V dataset version.

	Happiness	Fear	Anger	Sadness	Disgust	Surprise	Accuracy
Happiness	4660	91	134	**160**	25	159	89.12
Fear	136	3858	91	**234**	47	27	87.82
Anger	175	104	4107	**212**	70	55	86.96
Sadness	**430**	162	235	4072	30	92	81.10
Disgust	51	80	**148**	33	3289	19	**90.86**
Surprise	**270**	62	101	93	18	2459	81.88

Table 15. Precision, recall and *F*-measure of SVM on F5 dataset version.

	Precision	Recall	*F*-measure
Happiness	0.787	**0.862**	0.823
Fear	0.844	0.814	0.829
Anger	0.76	0.788	0.774
Sadness	0.775	0.778	0.777
Disgust	**0.87**	0.847	**0.859**
Surprise	0.81	0.742	0.775
Average	0.807	0.806	0.806

SVM: support vector machine.

Table 16. Precision, recall and *F*-measure results of SVM on F5_V dataset version.

	Precision	Recall	*F*-measure
Happiness	0.814	0.891	0.851
Fear	0.885	0.878	0.882
Anger	0.853	0.87	0.861
Sadness	0.848	0.811	0.829
Disgust	**0.945**	**0.909**	**0.927**
Surprise	0.875	0.819	0.846
Average	0.865	0.864	0.864

SVM: support vector machine.

validating the raw dataset, two datasets were formed: raw and validated datasets. Then, we subjected these two datasets to two different stemming methods: F5 and the Zemberek, which resulted four dataset versions, F5, F5_V, Z and Z_V.

In the next step, we used MI for feature selection on these four dataset versions. After having readied the dataset versions for classification, we applied four classification algorithms, CNB, J48, RF and SVM, to compare the validation, stemming and feature selection effects on the dataset versions. In the classification process, we evaluated the performance in terms of accuracy, precision, recall and *F*-measure. The results showed that for cases in which the validated dataset versions are used as the training datasets, classification results are higher than those using the raw datasets by 5.7% of the F5 stemmed dataset versions ($p = 0.000000092$) and by 5.6% of the Z dataset versions ($p = 0.00000013$). The improvements in the average of the accuracy values showed that the validation process is the correct choice for eliminating misleading declarations of emotion definitions by the participants in the survey. However, F5 and Zemberek stemmers showed similar performances. Therefore, statistical analysis indicated no significant difference between these two stemming methods ($p = 0.65$). For feature selection, we can say that both significant term selection approaches, selecting the first n terms or using a threshold, yielded similar results, and there is no significant difference between them ($p = 0.22$). Based on the results we obtained, the SVM classifier yielded the highest performance value among the others ($p = 0.02$).

For future work, we plan to use the TREMO dataset with different classifiers using different feature selection approaches. In addition, we plan to find the optimum value for selecting the first n terms to obtain the highest accuracy

value. Another future task is to extract the most valuable terms for each emotion category and to use them in the process of lexicon-based approach. Then, we plan to compare the results obtained using this approach with the results achieved using classification algorithms.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

Note

1. http://demir.cs.deu.edu.tr/tremo

References

[1] Kılınç D, Özçift A, Bozyigit F et al. TTC-3600: a new benchmark dataset for Turkish text categorization. *J Inf Sci* 2015; 43: 174–185.
[2] Scherer KR and Wallbott HG. Evidence for universality and cultural variation of differential emotion response patterning. *J Pers Soc Psychol* 1994; 66(2): 310–328.
[3] Calvo RA and Kim SM. Emotions in text: dimensional and categorical models. *Comput Intell* 2013; 29(3): 527–543.
[4] Danisman T and Alpkocak A. Feeler: emotion classification of text using vector space model. In: Volume 2: Proceedings of the AISB 2008 symposium on affective language in human and machine, Aberdeen, Scotland, 1–4 April 2008, pp. 53–59. Scotland: The Society for the Study of Artificial Intelligence and Simulation of Behaviour.
[5] Giachanou A and Crestani F. Opinion retrieval in Twitter using stylistic variations. In: *Proceedings of the 31st annual ACM symposium on applied computing*, Pisa, 4–8 April 2016, pp. 1077–1079. New York: ACM Press.
[6] Luo Z, Osborne M and Wang T. An effective approach to tweets opinion retrieval. *World Wide Web* 2015; 18(3): 545–566.
[7] Go A, Bhayani R and Huang L. Twitter sentiment classification using distant supervision. Project Report CS224N, 2009. Stanford, CA: Stanford University, pp. 1–12.
[8] Mohammad S. Portable features for classifying emotional text. In: *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies*, Montréal, QC, Canada, 3–8 June 2012, pp. 587–591. Stroudsburg, PA: Association for Computational Linguistics.
[9] Strapparava C and Valitutti A. Wordnet-affect: an affective extension of WordNet. In: *Proceedings of the 4th international conference on language resources and evaluation (LREC)*, Lisbon, 26–28 May 2004, pp. 1083–1086. Paris: ELRA.
[10] Mohammad SM and Turney PD. Emotions evoked by common words and phrases: using mechanical Turk to create an emotion lexicon. In: *Proceedings of workshop on computational approaches to analysis and generation of emotion in text (CAAGET)*, Los Angeles, CA, 5 June 2010, pp. 26–34. Stroudsburg, PA: Association for Computational Linguistics.
[11] Mohammad SM and Turney PD. Crowdsourcing a word-emotion association lexicon. *Comput Intell* 2013; 29(3): 436–465.
[12] Strapparava C and Mihalcea R. SemEval-2007 task 14: affective text. In: *Proceedings of the 4th international workshop on semantic evaluations*, Prague, 23–24 June 2007, pp. 70–74. Stroudsburg, PA: Association for Computational Linguistics.
[13] Chaffar S and Inkpen D. Using a heterogeneous dataset for emotion analysis in text. In: *Proceedings of the 24th Canadian conference on advances in artificial intelligence*, St. John's, NL, Canada, 25–27 May 2011, pp. 62–67. Berlin; Heidelberg: Springer.
[14] Kouloumpis E, Wilson T and Moore J. Twitter sentiment analysis: the good the bad and the OMG! In: *Proceedings of the fifth international AAAI conference on weblogs and social media (ICWSM)*, Barcelona, 17–21 July 2011, pp. 538–541. Menlo Park, CA: AAAI.
[15] Yang C, Lin KH and Chen H. Emotion classification using web blog corpora. In: *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*, Fremont, CA, 2–5 November 2007, pp. 275–278. New York: IEEE.
[16] Boynukalin Z. *Emotion analysis of Turkish texts by using machine learning methods*. Master's Thesis, Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, 2012.
[17] Demirci S. *Emotion analysis on Turkish tweets*. Master's Thesis, Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, 2014.
[18] Ekman P. An argument for basic emotions. *Cognition Emotion* 1992; 6(3): 169–200.
[19] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20(1): 37–46.
[20] Can F, Kocberber S, Balcik E et al. Information retrieval on Turkish texts. *J Am Soc Inf Sci Tec* 2008; 59(3): 407–421.
[21] Akin AA and Akin MD. Zemberek: an open source NLP framework for Turkic languages. Available at: https://github.com/ahmetaa/zemberek-nlp (2007, accessed 1 March 2016).

[22] Quinlan JR. C4.5: programs for machine learning. *Mach Learn* 1993; 16(3): 235–240.
[23] Manning CD, Raghavan P and Schütze H. *An introduction to information retrieval book*. Cambridge: Cambridge University Press, 2009.
[24] Rennie JDM, Shih L, Teevan J et al. Tackling the poor assumptions of Naïve Bayes text classifiers. In: *Proceedings of the 20th international conference on machine learning*, Washington, DC, 21–24 August 2003, pp. 616–623. Palo Alto, CA: AAAI.
[25] Kılınç D, Yücalar F, Borandağ E et al. Multi-level reranking approach for bug localization. *Expert Syst* 2016; 33(3): 286–294.
[26] Xu B, Guo X, Ye Y et al. An improved random forest classifier for text categorization. *J Comput* 2012; 7(12): 2913–2920.
[27] Platt JC. *Sequential minimal optimization: a fast algorithm for training support vector machines*. Technical report no. MSR-TR-98-14. USA: Microsoft Research, 1998.
[28] Witten IH and Frank E. *Data mining: practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufman, 2005.