Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 1780 – 1793

doi:10.3906/elk-1808-189

A hybrid sentiment analysis method for Turkish

Buket ERŞAHİN[1*], Özlem AKTAŞ[2], Deniz KILINÇ[3], Mustafa ERŞAHİN[1]

[1]Department of Computer Engineering, Graduate School of Natural and Applied Sciences, Dokuz Eylül University, İzmir, Turkey

[2]Department of Computer Engineering, Faculty of Engineering, Dokuz Eylül University, İzmir, Turkey

[3]Department of Software Engineering, Faculty of Technology, Celal Bayar University, Manisa, Turkey

Received: 27.08.2018 • **Accepted/Published Online:** 22.01.2019 • **Final Version:** 15.05.2019

Abstract: This paper presents a hybrid methodology for Turkish sentiment analysis, which combines the lexicon-based and machine learning (ML)-based approaches. On the lexicon-based side, we use a sentiment dictionary that is extended with a synonyms lexicon. Besides this, we tackle the classification problem with three supervised classifiers, naive Bayes, support vector machines, and J48, on the ML side. Our hybrid methodology combines these two approaches by generating a new lexicon-based value according to our feature generation algorithm and feeds it as one of the features to machine learning classifiers. Despite the linguistic challenges caused by the morphological structure of Turkish, the experimental results show that it improves the accuracy by 7% on average.

Key words: Sentiment analysis, opinion mining, social media, natural language processing

1. Introduction

Sentiment analysis (SA) is a field of text classification that allows to determine people's opinions and attitude on different products, services, and topics. The increasing popularity of social media in recent years has led to the explosion of data on the Web. The activities of users of social networking and friendship sites (e.g., Facebook), blogging and microblogging sites (e.g., Twitter), content and media sharing sites (e.g., YouTube), and shopping sites (e.g., Amazon, AliExpress) generate huge amounts of data. As it is almost impossible to read and interpret all these data manually, SA is required to automate such an exhaustive process.

SA is studied at three granularity levels, which are document level, sentence level, and aspect level. At the document level, the whole text is considered as an atomic unit and is assigned to a positive, negative, or neutral class as a result. At the sentence level, a sentence is identified as objective or subjective (holding an opinion). If it is subjective, it is assigned to a class; otherwise, it is ignored. The sentence level and document level SA approaches cannot discover more than one sentiment in a sentence or document. Aspect-based SA can discover different sentiments in a text with their related targeted terms. It can identify opinion tuples, which consist of a target term, target attribute (aspect), and target sentiment [1].

Considering the literature, methods used for SA are divided into three categories as machine learning (ML)-based approaches, lexicon-based approaches, and hybrid approaches [2]. In ML-based approaches, some ML algorithms are applied to predict the sentiment. On the other hand, the lexicon-based SA approach relies on sentiment lexicons. Lexicons are classified as dictionary-based or corpus-based according to their type of

*Correspondence: buketoksuzoglu@iyte.edu.tr

resources to find the sentiment polarities. The dictionary-based approach starts with finding seed sentiment words and expands with the synonyms and antonyms of these words. The corpus-based approach also starts with sentiment seed words, like the dictionary-based approach, but it expands with a large corpus of opinion words in the same context. The hybrid approach is a combination of the ML-based approach and lexicon-based approach [3]. A common strategy used to study SA is to apply either ML-based or lexicon-based approaches. On the other hand, some studies try to apply both, but not together in a hybrid approach, and compare the results of them. Our approach combines ML-based methods with lexicon-based methods as a hybrid approach and improves the results of SA. As far as we know, no previous research has investigated a hybrid approach in Turkish.

In this paper, we present a hybrid method for Turkish SA that is tested using three different datasets of Movie, Hotel, and Twitter. The main contributions of this study are as follows:

- To the best of our knowledge, it is the first study proposing and testing a hybrid SA method in Turkish.
- The first comprehensive Turkish SA dictionary, SentiTurkNet (STN) [4], is expanded using the Automated Synonym Dictionary (ASDICT) [5].
- Lemmatization in natural language processing (NLP) is adapted for Turkish SA to preserve the positive and negative meaning of tokens.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss related work on SA. Section 3 presents lexicon expansion and the steps of the proposed methodology. In Section 4 we consider the experimental setup, such as datasets and evaluation metrics. In Section 5 we evaluate our hybrid approach by comparing the results with lexicon-based and ML-based approaches. Finally, in Section 6, we discuss the conclusion and future work.

2. Related work

The related work is categorized into three classes according to the approach as ML-based, lexicon-based, and hybrid approaches. In ML-based approaches, supervised techniques are mostly used. In [6], Chinese mobile reviews were used as a dataset. This work showed that the mobile reviews have 17 Chinese characters on average, which are shorter than other short texts such as microblogs with 45 words on average. Labeling is done using iTune scores. Naive Bayes (NB) and support vector machine (SVM) algorithms are used and the results show that NB is better than SVM. Vinodhini and Chandrasekaran [7] examined the effect of principal component analysis (PCA) on SVM and NB algorithms. The experiments were done on product reviews. The results were improved using PCA for feature reduction in both algorithms. Pang [8] used NB, SVM, and maximum entropy on a movie review dataset and showed that binary representation is better than frequency representation. The Opinion Corpus for Arabic (OCA) was proposed by Rushdi Saleh et al. [9]. The corpus consists of 250 positive and 250 negative movie reviews. Various experiments were conducted on this corpus with NB and SVM. They observed that the best result using SVM over the OCA improved on the best result obtained with the Pang corpus, using trigrams to generate the word vectors. Govindarajan [10] proposed a hybrid method coupling NB and a genetic algorithm (GA), and experiments on movie reviews. The results showed that the hybrid method performs better than only NB or GA. In [11], Arabic tweets with dialectical words were tested with NB and SVM. Two versions of the dataset were studied; one was Tweets with dialectical words and second was with dialectical words as translated. The accuracy of the dataset with translated dialectical words was 3% better.

Lexicon-based approaches need a lexicon, which is generated either from an existing dictionary or extracted opinion words from a corpus. According to the polarity values in the lexicon, the general sentiment of the document is predicted. Baloglu and Aktas [12] introduced an opinion-mining application, which creates movie scores from blog pages. They got the sentiment scores from SentiWordNet [13] and declared that they produced accurate results close to IMDB results. The document-based Sentiment Orientation System [14] uses WordNet [15] to identify synonyms and antonyms so it gives the summary of the total number of positive and negative documents. Negation is also handled in the system. That work classified the document as positive if the number of positive words is greater; otherwise, the polarity is negative, and if the number of positive and negative words is equal, it is classified as neutral. They experimented on movie reviews and obtained accuracy of 63%.

Hybrid approaches use lexicon-based and ML-based approaches in combination. The language processing operations are done before the learning of ML algorithms. In [16], it was shown that a hybrid method using NLP techniques, semantic rules, and fuzzy sets performed well on movie reviews and achieved accuracy of 76%. Ohana and Tierney [17] calculated the sentiment direction using SentiWordNet and then applied the SVM classifier. They presented the results of applying the SentiWordNet lexical resource to the SA of film reviews. Their approach involves positive and negative term scores to determine sentiment, and they presented an improvement by building a dataset of relevant features using SentiWordNet as a source. Then they applied ML algorithms. The results indicated that SentiWordNet can be used as an important resource for SA. They obtained the best accuracy of 69.35% with SentiWordNet scores used as features.

Most of the research in the SA field focuses on English. There are a few works on SA on Turkish. Akgul et al. [18] compared the results of lexicon-based and character-based n-gram models. They preprocessed their Twitter dataset and ran n-grams. As a result, the lexicon method obtained accuracy of 70% and the n-gram model 69%, respectively. Turkmenoglu and Tantug [19] conducted a comparison of lexicon-based and ML-based approaches. After some preprocessing on the Twitter and movie datasets, they obtained accuracy of 75.2% on Twitter and 79% on the movie dataset by the lexicon-based approach. On the other hand, the best accuracy results of the ML-based approach were 85% for the Twitter dataset by SVM and 89.5% with SVM and NB on the movie dataset. In [20], experiments were done on hotel reviews with NB, SVM, and random forest (RF) algorithms. The best results were obtained by RF with accuracy of 82.2%. In [21], experiments were conducted on a Twitter dataset with NB, center-based classifier, multilayer perceptron (MLP), and SVM. According to the results, the best performance was achieved with MLP and SVM with accuracy values of 86% and 81% on the movie review dataset, respectively. Yildirim et al. [22] experimented on Tweets in the telecommunication area. NLP was used such for normalization, stemming, and negation handling. Ternary classification was achieved with accuracy of 79% using SVM. In [23], a Twitter dataset was employed with some different classification algorithms: SVM, NB, multinomial naive Bayes (MNB), and kNN. The results showed that the best accuracy was achieved with MNB at 66.08%. In [24], studies were done on movie reviews using unsupervised learning techniques. They made use of SentiStrength [25] to classify the texts by translating them. The results showed that the accuracy was 76% for binary classification. Kaya et al. [26] observed the sentiment analysis of Turkish political news. They used four different classifiers: NB, maximum entropy (ME), SVM, and character-based n-gram models. Their experimental results showed that ME with the n-gram language model was more effective than SVM and NB. Accuracy of 76% was achieved in binary classification of political news. Boynukalin [27] studied emotion analysis on Turkish texts by using a ML-based approach. Four types of emotions, joy,

sadness, fear, and anger, were examined on her own dataset and an accuracy of 78% was achieved. Erogul [28] investigated the effect of part-of-speech (POS) tags, word unigrams and bigrams, and negation handling. NLP processing was done with Zemberek, obtaining accuracy of 85% on binary classification of Turkish movie reviews. Dehkharghani et al. [29] proposed and evaluated a SA system for Turkish. Their system used STN and NLP techniques such as dependency parsing. They also covered different levels of granularities as well as some linguistic issues such as conjunction and intensification. Their system was evaluated on Turkish movie reviews and the obtained accuracies ranged from 60% to 79% in ternary and binary classification.

3. Materials and methods

Figure 1 shows the proposed hybrid method that aims to improve the accuracy of ML algorithms for SA by feeding them with a new lexicon-based feature. We apply five main steps, which consist of data collection, preprocessing and lexicon expansion, feature extraction with lemmatization (M1), polarity-based feature generation (M2), and ML. Data collection and ML algorithms used are explained in Section 4. All other steps are presented in the following subsections.

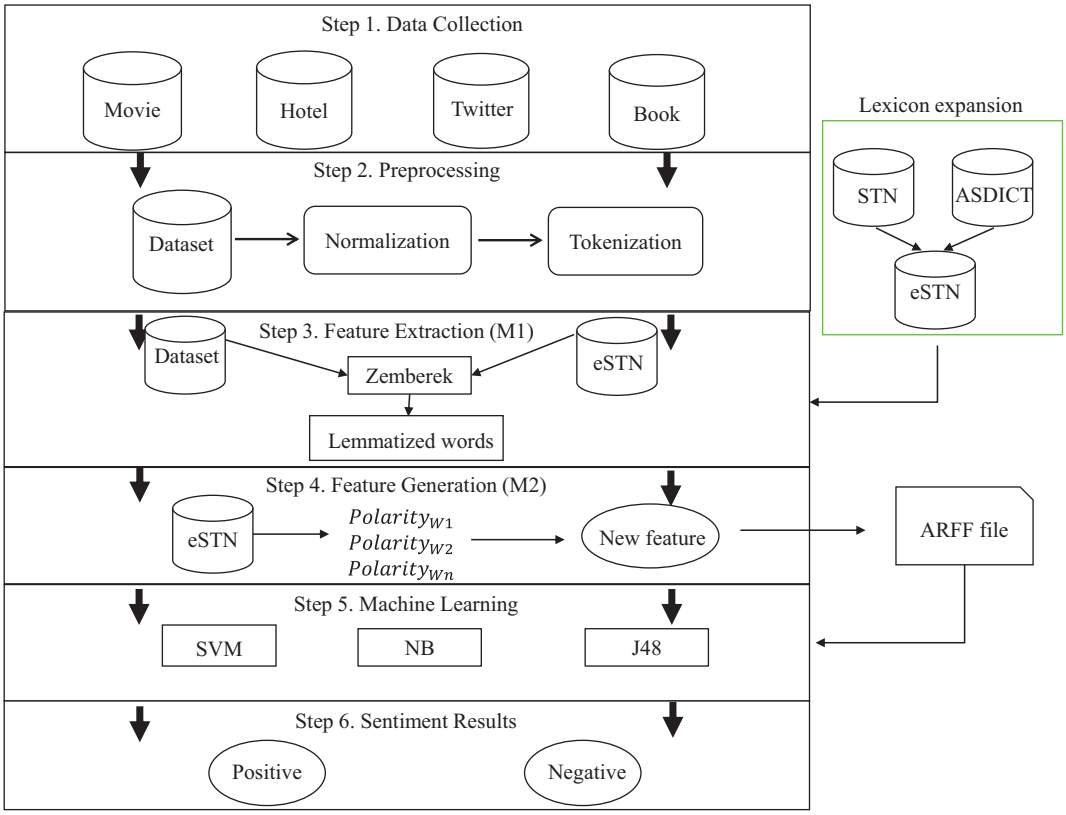

Figure 1. Our proposed framework.

3.1. Preprocessing

For a given dataset, the first step of SA is the preprocessing that involves a series of methods to improve the next phases. First, we normalize the input document utilizing the ITU NLP tool [30] and break it into tokens by using Zemberek. Then we lower the tokens to prevent mismatches because of case sensitivity. We also remove the tokens shorter than 2 characters.

3.2. Lexicon expansion

STN, the lexicon used in our study, is the first comprehensive polarity lexicon for Turkish and it is constructed using a semiautomatic approach. It is based on Turkish WordNet [31] and is mapped to both SentiWordNet and WordNet. It contains polarity values for all 15,000 synsets of Turkish WordNet, but the coverage size is small. In order to improve the performance of our matching process for lexicon-based feature generation, ASDICT is explored and utilized. The basic data source used in ASDICT is the Contemporary Turkish Dictionary (CTD), which includes more than 70,000 words and was published by Turkish Linguistic Association (Turkish abbreviation: TDK). To generate a reliable synonym dictionary and handle the ambiguities arising from the different meanings of words, supervised methods are used. For the synonym dictionary, all ambiguities are examined and finalized by the experts of the TDK and the College of Social Sciences and Literature of Dokuz Eylül University (DEU). In our lexicon expansion step, all words in ASDICT are searched in STN. If there is a match, the synonyms are added to STN with the polarity values that are already in STN. The new lexicon is called extended STN (eSTN). Objective terms are excluded from eSTN because binary classification is the goal. Multiword terms are also removed since our features are words as unigrams.

To evaluate the effectiveness of the expansion process, the coverage rates of STN and eSTN are compared on all datasets after applying the lemmatization. According to the results given in Table 1, the performance of eSTN varies depending on the type and size of the dataset. The average increase in the coverage rate is approximately 78%.

Table 1. Coverage rates of lexicons.

Coverage rates	STN	eSTN	Increase
Movie	449	685	53%
Hotel	415	780	88%
Tweet	145	284	96%

3.3. Feature extraction with lemmatization

After preprocessing, the datasets and eSTN are lemmatized using Zemberek. The aim of the lemmatization is to convert the word into a standard format by removing sentimentally insignificant suffixes. In this way the number of tokens is reduced. Lemmatization is done preserving negations in the word. For this, Turkish suffixes such as -me/-ma and -sız/-siz are conserved. The verbs are also translated into infinitive form, as seen in Table 2.

Table 2. Term lemmatization examples.

Term before lemmatization	Term after lemmatization
akılsız	akılsız
anlaşmazlık	anlaşmamak
beğenilmeyen	beğenmemek
dumanlı	dumanlı
gürültülü	gürültülü

The main challenge of text classification is dealing with a huge number of tokens. They prolong the learning time and affect the ML algorithms' performance negatively. To overcome this problem, feature extraction with our lemmatization approach is proposed. It is implemented by lemmatizing tokens of texts and eSTN terms and it also reduces the dimensionality, as seen in Table 3.

Table 3. Feature extraction with lemmatization.

Before lemmatization	Kesinlikle izlenip desteklenmesi gereken bir müthiş bir film konu olarak orjinal bir film olduğunu da söylemeliyim (16 tokens)
After lemmatization	kesin izlemek desteklemek gerek müthiş film konu olmak orjinal film olmak söylemek (12 tokens)

3.4. Polarity-based feature generation

One of the contributions of this paper is the generation of a new polarity-based feature, which improves the results significantly. In the feature extraction step, the tokens are lemmatized. In this step, the lemmatized tokens of a document are searched in eSTN and matching tokens are used to create the polarity-based feature. Using eSTN, the number of positive tokens and the number of negative tokens are calculated, and the value of the new feature is calculated considering the algorithm in Figure 2.

- **Input:** S_1 - document as String
- **Output:** polarity_prediction – predicted sentiment class (new feature)
-
- **1:** **Begin**
- **2:** $polarity\_score_{S1} \leftarrow 0$ //initialize polarity score
- **3:** **For** i←0,., numberOfTokens **do**
- **4:** **If** (S_1[i] is positive) **Then** //as a result of STN matching
- **5:** $pos_{S1} \leftarrow pos_{S1} + 1$ //number of positive tokens
- **6:** **Else If** (S_1[i] is negative) **Then**
- **7:** $neg_{S1} \leftarrow neg_{S1} + 1$ //number of negative tokens
- **8:** $polarity\_score_{S1} \leftarrow polarity\_score_{S1} - polarity_{S1[i]}$
- **9:** **If** ($pos_{S1} - neg_{S1} >= 2$) **Then**
- **10:** $polarity\_prediction_{S1} \leftarrow$ pos
- **11:** **Else If** ($neg_{S1} - pos_{S1} >= 2$) **Then**
- **12:** $polarity\_prediction_{S1} \leftarrow$ neg
- **13:** **Else**
- **14:** **If** ($polarity\_score_{S1} < 0$) **Then**
- **15:** $polarity\_prediction_{S1} \leftarrow$ neg
- **16:** **Else if** ($polarity\_score_{S1}$) > 0 **Then**
- **17:** $polarity\_prediction_{S1} \leftarrow$ pos
- **18:** **Else**
- **19:** $polarity\_score_{S1} \leftarrow$ neut
- **20:** **return** $polarity\_prediction_{S1}$
- **21:** **End**

Figure 2. Feature generation algorithm.

As seen in Figure 3, the proposed feature generation algorithm takes the text as input and creates the lexicon-based new feature as output. After preprocessing and feature extraction, the selected features are "harika", "süper değil", "güzel", "eski", and "iyi" for the given example text. As was mentioned before, when "değil" is encountered, the polarity value of token just before it is negated. This means the values of negative and polarity scores are interchanged, as in Table 4. Then the polarity values and class labels of the tokens are taken from eSTN and processed according to our proposed algorithm. Based on the results of the algorithm, the

number of positive tokens, the number of negative tokens, the difference between them, total positive polarity, total negative polarity, and difference between them are calculated. Since the difference between positive tokens and negative tokens is not greater than or equal to 2 in this example text, the difference between positive polarity values and negative polarity values is calculated. It is found as 0.66, and since it is a positive value, a new feature is generated as positive.

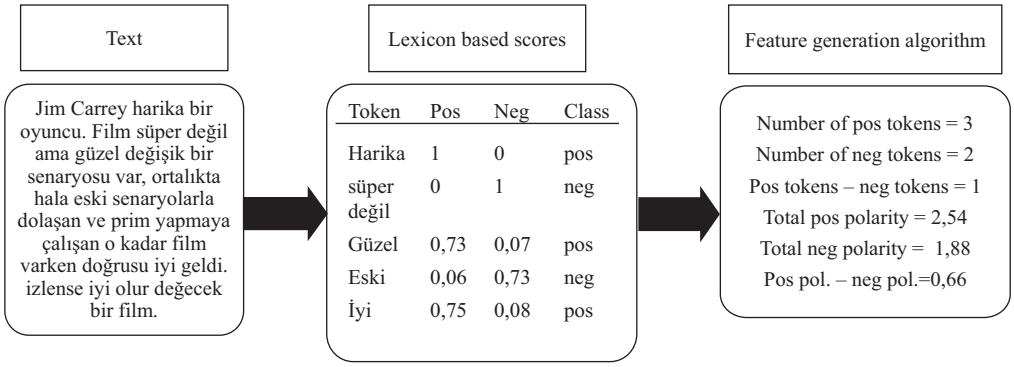

Figure 3. Feature generation scenario.

Table 4. Handling negation.

Term	Positive polarity	Neutral polarity	Negative polarity
güzel (beautiful)	1	0	0
güzel değil (not beautiful)	0	0	1
fena (bad)	0.035	0.02	0.945
fena değil (not bad)	0.945	0.02	0.035

The threshold value in this algorithm is selected with a grid search [32]. It is a technique that scans the data to configure the optimal parameters for a given model and works in an iterative way. In our model, we experiment with parameters 1 to 5. The grid search iterates through each of them and compares the result for each value. It finds the best parameter as 2 for our model.

4. Experimental setup

4.1. Datasets

In this study, experiments are done by using three different datasets to evaluate the results of the methodology on different types of data. Movie review and hotel review datasets are downloaded from the Hacettepe University Multimedia Information Retrieval Group's website [33]. Movie reviews on this website are collected from beyazperde.com and hotel reviews are collected from otelpuan.com. All extracted movie reviews are rated by their own authors according to stars. One or 2 stars is classified as negative, while 4 or 5 stars is classified as positive. In the similar way, hotel reviews are rated between 0 and 100 instead of stars. The negative reviews are selected from 0 to 40 point reviews and the positive from 80 to 100 point reviews. A completely different dataset consisting of Tweets is also used in the experiments to control the accuracy of the proposed methodology. This dataset is taken from the website of the Kemik NLP group of Yıldız Technical University. It consists of 3000

Turkish tweets having three classes for SA. The statistics of the datasets including the number of instances, sentences, and tokens are represented in Table 5.

Table 5. Statistics of the datasets.

Datasets	# of instances	# of sentences	# of tokens
Movie	49,476	106,813	1,345,726
Hotel	11,164	17,874	738,216
Tweets	1756	2535	19,056

4.2. ML algorithms and evaluation metrics

The algorithms used in the study are NB, SVM, and J48. NB is selected as a probabilistic classifier, SVM is selected as a linear classifier, and J48 is selected as a decision tree classifier. NB is one of the simplest and most commonly used machine learning algorithms used for text classification and based on the statistical Bayes theorem and conditional probability. The NB classifier presumes that the impact of the value of a feature on a given class is independent of the values of other attributes. SVMs are based on the structural risk minimization principle [34], which is the idea of finding a hypothesis (h) with the lowest error [35]. The error is the probability that h will have when it encounters new or randomly selected data. They can learn independently of the dimensionality of features and therefore work well for text categorization. J48 is a C4.5 decision tree algorithm for classification, based on binary trees. The main idea is to divide the data into ranges based on the attribute values in the training set [36].

The evaluation metrics used are accuracy, precision, recall, and f-measure, which are defined using the terms in Table 6.

Table 6. Definition of confusion matrix.

		Predicted class	
		P	N
Actual class	P	TP (True positives): The number of true positives, i.e. the number of files that are classified as positive correctly	FN (False negatives): The number of false negatives, i.e. the number of files that are classified as negative incorrectly
	N	FP (False positives): The number of false positives, i.e. the number of files that are classified as positive incorrectly	TN (True negatives): The number of true negatives, i.e. the number of files that are classified as negative correctly

Accuracy (Acc) is the ratio of the number of documents that are correctly classified to the total number of documents. The calculation of accuracy is given in Eq. (1).

$$Acc = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

Precision (Pr) is the probability that a randomly selected document is retrieved as relevant. It is calculated as the ratio of the total number of positive files that are correctly classified to the total number of positive classified files, as in Eq. (2):

$$Pr = TP/(TP + FP) \tag{2}$$

Recall (Re) is the probability that a randomly selected relevant document is retrieved in a search. It is calculated as the ratio of total number of positive files that are correctly classified to the number of positive files that are in the dataset, as in Eq. (3):

$$Re = TP/(TP + FN) \tag{3}$$

The F-measure (Fm) is the harmonic mean of precision and recall and it is calculated as in Eq. (4):

$$Fm = 2 * Pr * Re/(Pr + Re) \tag{4}$$

5. Experimental results

All datasets used in the experiments are balanced and have separate training and test sets, except the Twitter dataset, and it is run with 10-fold cross-validation. According to the experimental results, there are improvements in all of the three datasets. The results show that our hybrid approach outperforms both the lexicon-based and ML-based results in all datasets as seen in Table 7.

Table 7. Summary of experimental results.

Classifier	Dataset	Method	Average			Acc.
			Pr	Re	Fm	
Movie	NB	ML	0.83	0.804	0.8	80.35%
		Hybrid	0.891	0.889	0.889	88.93%
	SVM	ML	0.799	0.799	0.798	79.85%
		Hybrid	0.863	0.863	0.863	86.31%
	J48	ML	0.689	0.674	0.667	67.35%
		Hybrid	0.781	0.779	0.779	77.92%
	Lexicon		0.67	0.79	0.725	70.93%
Hotel	NB	ML	0.875	0.838	0.834	83.80%
		Hybrid	0.909	0.9	0.899	89.98%
	SVM	ML	0.912	0.911	0.911	91.14%
		Hybrid	0.92	0.92	0.92	91.96%
	J48	ML	0.869	0.861	0.86	86.10%
		Hybrid	0.892	0.89	0.889	88.96%
	Lexicon		0.73	0.91	0.81	78.88%
Twitter	NB	ML	0.7	0.702	0.701	70.21%
		Hybrid	0.834	0.834	0.834	83.37%
	SVM	ML	0.716	0.708	0.71	70.84%
		Hybrid	0.822	0.818	0.819	81.83%
	J48	ML	0.672	0.667	0.647	66.69%
		Hybrid	0.729	0.727	0.728	72.72%
	Lexicon		0.53	0.81	0.64	62.81%

To check the effectiveness of the new feature, the attributes are ranked using a filter-based attribute selection method, with information gain (IG) as an attribute evaluator and ranker as a search method, then sorted according to IG score. The experimental results are shown in Table 8. It is clearly seen that our new

attribute named "type" is the first ranked attribute, having by far the best IG ranking score in all three datasets. The scores are 0.17388 in Movie, 0.32817 in Hotel, and 0.04737 in the Twitter dataset, respectively. The score in the Twitter dataset is less than the others because the Tweets in the dataset are very short and there are some abbreviations and jargon, which makes finding strong sentiment words harder. Despite this, our new feature is still in the first rank. Although the second and third ranked features are the most used and powerful sentiment words in the language, the new feature has more impact in terms of sentiment.

Table 8. IG scores of new generated feature; best results in bold font.

Movie dataset			Hotel dataset			Twitter dataset		
Id	Name	Score	Id	Name	Score	Id	Name	Score
4200	**type**	**0.17388**	**3053**	**type**	**0.32817**	**2468**	**type**	**0.04737**
103	kötü (bad)	0.03646	1251	berbat (terrible)	0.15113	68	güzel (beautiful)	0.03886
26	harika (wonderful)	0.03189	19	güzel (beautiful)	0.12453	66	hayat (life)	0.03885

To improve the generalizability of the results, they are tested using three different algorithms, i.e. NB as a probabilistic classifier, SVM as a linear classifier, and J48 as a decision tree classifier. As a result of nine runs with three algorithms, the minimum difference between baseline and our approach's accuracy was 1.12% in the Hotel dataset with SVM. On the other hand, the maximum difference was 13.33% in the Twitter dataset with NB, as seen in Figure 4. The average improvement in all datasets with all algorithms was 7%.

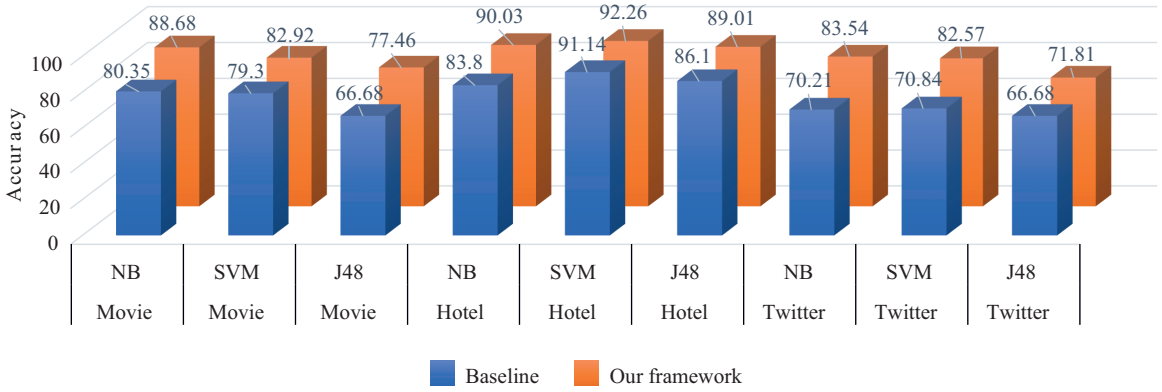

Figure 4. The experimental results of different ML algorithms.

In order to evaluate the statistical significance of the results, we have performed a two-way ANOVA test. In Figure 5, test results in terms of accuracy values are presented. The statistical test results can also be examined in Table 9. In this table, DF, SS, MS, and F denote degrees of freedom, adjusted sum of squares, mean squares, F-statistics, and probability value, respectively. As can be observed from the results, there is statistically significant difference ($P < 0.001$) for the means of the compared classifiers, datasets, and methods.

In addition, the 95% confidence interval for the compared algorithms based on the pooled standard deviation is presented in Figure 5, which supports the results presented in Table 9. Based on the statistical significances between the empirical results, Figure 5 is divided into two regions denoted by red dashed lines.

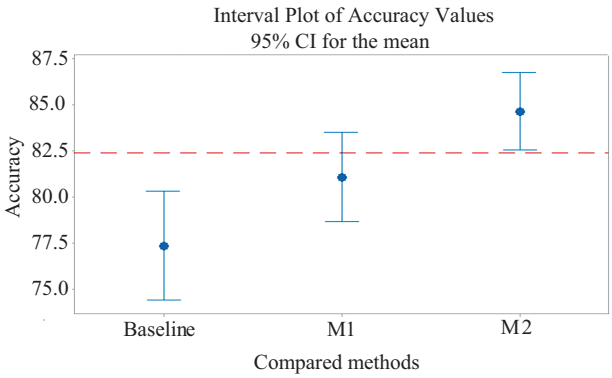

Figure 5. Interval plot of accuracy.

Hence, Figure 5 indicates that the differences between the results obtained by the proposed scheme (M2) are statistically significant compared to the results obtained by the baseline methods.

Table 9. Results of ANOVA test.

Analysis of variance (for F-measure)					
Source		DF	Adj SS	Adj MS	F-value
	Classifier	2	0.02483	0.012413	13.06
	Dataset	2	0.09837	0.049184	51.77
	Methods	2	0.02664	0.013320	14.02
Error		20	0.01900	0.000950	
Total		26	0.16883		
Analysis of variance (for accuracy)					
Source		DF	Adj SS	Adj MS	F-Value
	Classifier	2	225.3	112.649	12.44
	Dataset	2	969.6	484.818	53.56
	Methods	2	239.5	119.739	13.23
Error		20	181.0	9.052	
Total		26	1615.5		

There are significant improvements achieved with our hybrid SA framework in Turkish in all runs. SVM usually has the highest accuracy of all classification algorithms due to its robust nature, but it requires a large training set and very long training time. The NB method is improved with our approach and surpassed the SVM and J48 in all cases except the Hotel dataset.

Finally, we compare our approach to previous SA studies using the same datasets. These studies, their techniques, and accuracy values are given in Table 10. First, in [37], the authors investigated the feasibility of active learning for Turkish SA. The aim of active learning is to get the same or better results with smaller amounts of training data. They experimented with the Twitter dataset that we used and the NB method. The results of the system with active learning were better than only NB with accuracy values 64% and 62.6%, respectively. Another study [38] using the same Twitter dataset compared the performance of four feature selection methods using logistic regression. They showed that query expansion ranking (QER) and ant colony

optimization (ACO) methods outperformed other traditional feature selection methods for SA. They evaluated their results with Fm using 5-fold CV and got the best results with QER. Movie and hotel datasets were prepared and used in [33]. They proposed an automatic translation approach to create a lexicon for a new language. They used English resources mapping automatically to Turkish and constructed three different lexicons using different methods. Finally, they experimented with their lexicons and got the best accuracy value of 70.35% for Movie and 80.68% for Hotel utilizing TSDp, which is a lexicon prepared by parallel-based translation approach. Their ML-based results with SVM were 84.6% and 79.7% in the Movie and Hotel datasets, respectively. By all accounts, our hybrid method performs better on all the same datasets.

Table 10. Comparison of studies.

Method	Dataset	Technique	Results
[37]	Twitter	NB	Acc: 62.6%
[37]	Twitter	NB + active learning	Acc: 64%
[38]	Twitter	Logistic regression + QER	Fm: 0.779
[33]	Movie	Lexicon	Acc: 70.35%
[33]	Movie	SVM	Acc: 84.6%
[33]	Hotel	Lexicon	Acc: 80.68%
[33]	Hotel	SVM	Acc: 79.7%
Our method	Twitter	Hybrid (NB+eSTN)	Acc: 83.37%
Our method	Hotel	Hybrid (SVM+eSTN)	Acc: 91.96%
Our method	Movie	Hybrid (SVM+ eSTN)	Acc: 86.31%
Our method	Movie	Lexicon	Acc: 70.93%
Our method	Hotel	Lexicon	Acc: 78.88%

6. Conclusion and future work

In this study, we present a hybrid approach for SA in Turkish and test it with three different datasets (Movie, Hotel, and Twitter) by three different ML algorithms of NB, SVM, and J48. We show that the accuracy of the SA for all datasets can be improved by combining the ML-based and lexicon-based approaches. On the lexicon-based side, STN is expanded with ASDICT and a lexicon score is calculated based on the polarity of the words in eSTN. On the ML-based side, three different ML models are run by feeding the generated feature as one of the features. The ranking of all features based on the IG scores show that the lexicon-based new feature is at the top of the list, confirming its relevance. Additionally, the results show that our hybrid approach outperforms the other two approaches. To the best of our knowledge, this is the first study on a hybrid SA framework for Turkish.

One of the future directions for the proposed framework consists of research on aspect-based SA and use of its subtasks to improve the overall performance of the system. As another future work we would like to evaluate the proposed method on some English datasets to check its effectiveness in multilingual environments. Finally, word vectors such as Word2Vec may be used to improve the quality of the feature selection process.

References

[1] Boudad N, Faizi R, Rachid OHT, Chiheb R. Sentiment analysis in Arabic: a review of the literature. Ain Shams Engineering Journal 2018; 9 (4): 2479-2490. doi: 10.1016/j.asej.2017.04.007

[2] Maynard D, Funk A. Automatic detection of political opinions in tweets. In: Proceedings of the 8th International Conference on the Semantic Web; Heraklion, Greece; 2012. pp. 88-99.

[3] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Engineering Journal 2014; 5 (4): 1093-1113. doi: 10.1016/j.asej.2014.04.011

[4] Dehkharghani R, Saygin Y, Yanikoglu B, Oflazer K. SentiTurkNet: A Turkish polarity lexicon for sentiment analysis. Language Resources and Evaluation 2015; 50 (3): 667-685. doi: 10.1007/s10579-015-9307-6

[5] Aktaş Ö, Birant Ç, Aksu B, Çebi Y. Automated synonym dictionary generation tool for Turkish (ASDICT). BILIG - Turk Dunyasi Sosyal Bilimler Dergisi 2013; 65 (9): 47-68

[6] Zhang L, Hua K, Wang H, Qian G, Zhang L. Sentiment analysis on reviews of mobile users. Procedia Computer Science 2014; 34 (11): 458-465. doi: 10.1016/j.procs.2014.07.013

[7] Vinodhini G, Chandrasekaran RM. Effect of feature reduction in sentiment analysis of online reviews. International Journal of Advanced Research in Computer Engineering & Technology 2013; 2 (6): 2278–1323

[8] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques; In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA; 2002. pp. 79-86.

[9] Rushdi-Saleh M, Martín-Valdivia MT, Ureña-López LA, Perea-Ortega JM. OCA: Opinion corpus for Arabic. Journal of the Association for Information Science and Technology 2011; 62 (10): 2045–2054

[10] Govindarajan M. Sentiment analysis of movie reviews using hybrid method of naive Bayes and genetic algorithm. International Journal of Advanced Computer Research 2013; 3 (4): 139-146

[11] Duwairi RM. Sentiment analysis for dialectical Arabic. In: Proceedings 6th International Conference on Information and Communication Systems; Amman, Jordan; 2015. pp. 166-170.

[12] Baloglu A, Aktas MS. An automated framework for mining reviews from blogosphere. International Journal of Advances in Internet Technology 2010; 3 (4): 234-244.

[13] Esuli A, Sebastiani F. SENTIWORDNET: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation; Genoa, Italy; 2006. pp. 417-422.

[14] Sharma R, Nigam S, Jain R. Opinion mining of movie reviews at document level. International Journal on Information Theory 2014; 3 (3): 13-21. doi: 10.5121/ijit.2014.3302

[15] Fellbaum C. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). Cambridge, MA: MIT Press, 1998.

[16] Appel O, Chiclana F, Carter J, Fujita H. A hybrid approach to sentiment analysis. In: IEEE Congress on Evolutionary Computation; Sendai, Japan; 2016. pp. 4950-4957.

[17] Ohana B, Tierney B. Sentiment classification of reviews using SentiWordNet. In: 9th IT&T Conference; Dublin, Ireland; 2009; pp. 10-19. doi: 10.21427/D77S56

[18] Akgül ES, Ertano C, Diri B. Sentiment analysis with Twitter. Pamukkale University Journal of Engineering Sciences 2016; 22 (2): 106-110. doi: 10.5505/pajes.2015.37268

[19] Turkmenoglu C, Tantug AC. Sentiment analysis in Turkish media. In: International Conference on Machine Learning; Beijing, China; 2014; pp. 32-42. doi: 10.13140/2.1.1502.1125

[20] Oğul BB, Ercan G. Sentiment classification on Turkish hotel reviews. In: 24th Signal Processing and Communication Application Conference; Zonguldak, Turkey; 2016. pp. 497-500.

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

[22] Yildirim E, Çetin F, Eryigit G, Temel T. The impact of NLP on Turkish sentiment analysis. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 2017; 7 (1): 43-51.

[23] Çoban Ö, Özyer B, Özyer GT. Sentiment analysis for Turkish Twitter feeds. In: 23nd Signal Processing and Communications Applications Conference; Malatya, Turkey; 2015. pp. 2388-2391.

[24] Vural AG, Cambazoglu BB, Senkul P, Tokgoz ZO. A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish. In: Computer and Information Sciences III; London, UK; 2012. pp. 437-445.

[25] Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 2012; 63 (1): 163-173. doi: 10.1002/asi.21662

[26] Kaya M, Fidan G, Toroslu IH. Sentiment analysis of Turkish political news. In: International Conferences on Web Intelligence and Intelligent Agent Technology; Washington, DC, USA; 2012. pp. 174-180.

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

[28] Eroğul U. Sentiment analysis in Turkish. MSc, Middle East Technical University, Ankara, Turkey, 2012.

[29] Dehkharghani R, Yanikoglu B, Saygin Y, Oflazer K. Sentiment analysis in Turkish at different granularity levels. Natural Language Engineering 2017; 23 (4): 535-559. doi: 10.1017/S1351324916000309

[30] Eryiğit G, Torunoğlu-Selamet D. Social media text normalization for Turkish. Natural Language Engineering 2017; 23 (6): 1-41. doi: 10.1017/S1351324917000134

[31] Ehsani R, Solak E, Yildiz OT. Constructing a WordNet for Turkish using manual and automatic annotation. ACM Transactions on Asian Language Information Processing 2018; 17 (3): 1-15. doi: 10.1145/3185664

[32] Thisted RA. Elements of Statistical Computing: Numerical Computation. New York, NY, USA: Chapman & Hall, 1988.

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

[34] Vapnik VN. The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 1995.

[35] Joachims T. Text categorization with support vector machines: learning with many relevant features. In: European Conference on Machine Learning; Chemnitz, Germany; 1998. pp. 137-142.

[36] Goyal A, Mehta R. Performance comparison of naïve Bayes and J48 classification algorithms. International Journal of Applied Engineering Research 2012; 7 (11): 1389-1393

[37] Cetin M, Fatih AM. Active learning for Turkish sentiment analysis. In: IEEE International Symposium on Innovations in Intelligent Systems and Applications; Turkey; 2013. pp. 1-4.

[38] Parlar T, Sarac E, Ozel SA. Comparison of feature selection methods for sentiment analysis on Turkish Twitter data. In: 25th Signal Processing and Communications Applications Conference; Turkey; 2017. pp. 1-4.