Research Article
Journal of Intelligent Systems: Theory and Applications 4(2) (2021) 103-112
DOI: 10.38016/jista.854250

Analysis of Turkish Sentiment Expressions About Touristic Sites Using Machine Learning

Muhammed Çağrı Aksu[1]*, Ersin Karaman[2]

[1]Artvin Çoruh University, Rectorate, Department of Informatics, Artvin, Turkey

[2] Ankara Hacı Bayram Veli University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Ankara, Turkey

cagriaksu@artvin.edu.tr, ersin.karaman@hbv.edu.tr

Abstract
Analyzing data by inferring from unstructured data about customers is one of the main purposes of the tourism and many other industries as well. However, performing unstructured data analysis using traditional methods is quite inconvenient and costly. This can be overcome by using sentiment analysis, an area of application of text mining. Since there is no proven methodology for sentiment analysis, researchers often perform their studies by trial and error. Many studies on sentiment analysis have focused on comparing the preprocessing or the performance of various machine learning algorithms. Both for these reasons and since research on sentiment analysis with Turkish content is limited, this study aimed to determine the effects of labeling, stemming, and negation on the success of sentiment analysis using Turkish touristic site analysis. From the data set prepared for this study, 12 different variations were created according to labeling, number of classes, stemming, and negation. These data sets were classified using the algorithms Naive Bayes (NB), Multinominal Naive Bayes (MNB), k-Nearest Neighbor, and Support Vector Machines (SVM), often used in sentiment analyses, and the findings were compared.

Keywords: Text mining, sentiment analysis, machine learning, unstructured data analysis, classification, naive bayes, multinominal naive bayes, support vector machines

Turistik Mekanlar Hakkındaki Türkçe Duygu İfadelerinin Makine Öğrenmesi Yöntemleri ile İncelenmesi

Öz
Müşteriler ile ilgili yapılandırılmamış verilerden çıkarımlar yaparak bu verileri analiz etmek birçok sektör için olduğu gibi turizm sektörü için de temel amaçlardandır. Yapılandırılmamış veri analizinin geleneksel yöntemlerle gerçekleştirilmesi oldukça zahmetli ve maliyetli olmaktadır. Metin analizi uygulama alanlarından biri olan duygu analizi kullanılarak bu sorunun üstesinden gelinebilmektedir. Duygu analizi çalışmalarında henüz kanıtlanmış bir metodoloji bulunmadığı için araştırmacılar genellikle deneme yanılma yoluyla çalışmalarını yürütmektedirler. Duygu analizi alanında yapılan birçok çalışma duygu analizi ön işlemlerinin ya da farklı makine öğrenimi algoritmalarının performanslarının karşılaştırılması üzerinedir. Hem bu nedenlerden dolayı hem de Türkçe içeriklerle gerçekleştirilmiş duygu analizi çalışmalarının kısıtlı olmasından dolayısıyla bu çalışmada Türkçe turistik mekân incelemeleri kullanılarak duygu analizi ön işlemlerinden etiketleme, köklerine ayırma ve olumsuzlaştırma işlemlerinin duygu analizinin başarısına olan etkileri tespit edilmeye çalışılmıştır. Bu nedenle bu çalışma için hazırlanan veri setinden etiketlenme şekline, sınıf sayısına, köklerine ayırma ve olumsuzlaştırma durumlarına göre 12 farklı varyasyon oluşturulmuştur. Oluşturulan bu veri setleri duygu analizi çalışmalarında sıklıkla kullanılan Naive Bayes (NB), Multinominal Naive Bayes (MNB), k-Nearest Neighbor ve Support Vector Machines (SVM) algoritmalarıyla sınıflandırılarak elde edilen sonuçlar karşılaştırılmıştır.

Anahtar Kelimeler: Metin madenciliği, duygu analizi, makine öğrenmesi, yapılandırılmamış veri analizi, sınıflandırma, naive bayes, multinominal naive bayes, destek vektör makineleri

* Corresponding Author.
E-mail: cagriaksu@artvin.edu.tr
** This study is derived from the first author's doctoral thesis.

Recieved : 05 Jan 2021
Revision : 09 Mar 2021
Accepted : 16 Apr 2021

1. Introduction

Structured data obtained by traditional methods are insufficient for dynamically analyzing changing customer trends (Esen & Türkay, 2017). Therefore, new mechanisms have emerged that allow the dynamic analysis of changing customer trends and support the decision-making of managers in organizations. These mechanisms enable data-driven decision-making using unstructured data sources such as social media posts, blogs, and web server logs (Provost & Fawcett, 2013). Most of the data used in organizations consist of unstructured data, with estimations of around 80% (Beal; Blumberg & Atre, 2003; Lohr, 2012). Therefore, it is of great importance to analyze unstructured data and to extract meaningful information that will be useful for the organization. Through sentiment analysis, which is one of the sub-research areas of text mining, people's opinions, evaluations, attitudes, and sentiments about products, services, and activities can be analyzed (B. Liu, 2012). In other words, unstructured data can be analyzed using sentiment analysis.

Sentiment analysis can be performed in two ways, namely, dictionary-based method and machine learning method (Can & Alataş, 2017). In the dictionary-based method, there is a dictionary called sentiment dictionary that contains a large number of words, and each word has a sentimental polarity, that is, the word's positivity, negativity, or neutrality degree. In this method, the sentimental polarity of the text is calculated by searching each word, the sentimental state of which is to be determined, in the sentiment dictionary (Baccianella et al., 2010; Esuli & Sebastiani, 2006; Kuvd., 2006). Then, data of unknown classes are classified using this model (Pangvd., 2002). In some sentiment analysis studies conducted with machine learning, the classes of the data are determined manually by the researchers (Kulcu & Dogdu, 2016; K.-L. Liu, Li, & Guo, 2012). Some studies perform a labeling process using metadata. These metadata are often scores ranging from one to five or the emojis in the text to be analyzed (Chang, Ku, & Chen, 2019; Gezici & Yanıkoğlu, 2018; Taecharungroj & Mathayomchan, 2019; Türkmenoglu & Tantug, 2014).

Aydoğan and Akcayol (2016), Özyurt and Akçayol (2018), Can and Alataş (2017) reviewed recent studies on sentiment analysis based on their methods, areas of application, and data sets. Considering these reviews and other research in the literature (Bilgin & Şentürk, 2017; Çoban et al., 2015; Kaynar et al., 2016; Kızılkaya, 2018; Meral & Diri, 2014; Salur et al., 2019; Toçoğlu, 2018; Türkmenoğlu, 2015), it is observed that the studies on sentiment analysis are often focused on the classification of examinations, tweets, or comments in a certain area using various machine learning techniques and the comparison of these classification results. It is also found that sentiment analysis studies in the context of tourism are quite limited and there are suggestions to conduct further analysis including Turkish content. Thus, here, it was aimed to create a data set that allows the sentiment analysis of texts about Turkish touristic sites, to determine the effect of labeling on the success of the classification, and to examine the effect of tokenization, stemming, and negation on the success of classification.

To achieve the aims of the study, it is first necessary to create a data set. To create a data set suitable for the subject of the current sentiment analysis, reviews on Tripadvisor, Google Maps, and Foursquare, websites that are frequently mentioned in research on tourism and that allow tourists to make comments or reviews about the places they visit, were used. From these sites, reviews by tourists for 203 touristic sites in the cities of Trabzon, Artvin, and Rize in the Eastern Black Sea region of Turkey were obtained. A total of 49031 reviews were obtained on 10.03.2020. This data consisted of reviews by tourists, the date of their reviews, and scores ranging from one to five given by the tourists. Using these data, the data sets described in the third section were created. Using each of these data sets, machine learning models were created using the classic machine learning algorithms NB, MNB, k-NN, and SVM. The machine learning models were validated by 10-repetitive cross-validation and the findings were obtained by the f-scores of the models.

The second section describes sentiment analysis, the method of the study. The third section explains how the experiments were performed. The fourth section demonstrates the findings. Finally, the last section discusses these findings.

2. Materials and Methods

This section briefly explains how the data are obtained, the aforementioned machine learning algorithms, and the method, sentiment analysis.

2.1. Data Collection

As stated, the data was obtained from the websites of Tripadvisor, Google Maps, and Foursquare to create the data sets. First, a list of touristic sites in the cities of Trabzon, Artvin, and Rize in the Eastern Black Sea region of Turkey was obtained from karadeniz.gov.tr. Then, these touristic sites were searched on the websites and the pages containing comments or reviews on them were obtained. The links to these pages were sent to the

Datashake[1] data scraping[2] service with a php script. The reviews were then recorded in the database. Finally, the data in the database were applied the processes specified in the experiment section and the dataset(s) was created.

2.2. Sentiment Analysis

The process of automatically discovering some previously unknown information from different written sources is called text mining (Hearst, 2003). Text mining is divided into seven areas of application as text clustering, text classification, web mining, information extraction, natural language processing, concept extraction, and information retrieval and each area of application has its specific features (Miner et al., 2012). Sentiment analysis is located at the intersection of opinion extraction and document classification (Miner et al., 2012). Sentiment analysis, also known as opinion mining, is a field of study that analyzes people's views, evaluations, attitudes, and feelings about assets such as products, services, organizations, individuals, activities, topics, and their characteristics (B. Liu, 2012). Sentiment analysis is often done by classifying the sentiment in a text in a binary (positive-negative) or ternary (positive-negative-neutral) form (Şeker, 2016). Sentiment analysis can be performed in two ways: dictionary-based method and machine learning method (Can & Alataş, 2017). In the dictionary-based method, each word in the text is searched in dictionaries with predetermined polarities, the opinion score of the text is calculated, and classification is performed (Baccianella, Esuli, & Sebastiani, 2010; Esuli & Sebastiani, 2006; Ku, Liang, & Chen, 2006). In the machine learning method, a machine learning model with labeled data is created and data with unknown classes are tried to be classified using the created model (Pang, Lee, & Vaithyanathan, 2002). In this study, sentiment analysis was performed using the machine learning method, as it has been stated to be the superior method in the literature (Özyurt & Akçayol, 2018).

Since many classification algorithms used in sentiment analysis by machine learning cannot work with categorical data, they must be converted into numerical data. In sentiment analysis, the conversion of categorical data is often carried out by the bag-of-word (BOW) method ("bag-of-word model," 2007; Harris, 1954). In this method, every single term in the text (a word, a sentence, or a certain number of characters) is considered an attribute and the frequency of each term in the text is assigned as the value of the attribute. Thus, categorical text data is converted into digital form.

In many sentiment analysis studies, the data is preprocessed, consisting of steps such as tokenization, normalization, stemming, stop words removal, and term weighting to increase classification success and to reduce the attribute size (Aydoğan&Akcayol, 2016; Çoban, 2016; Çoban et al., 2015; Meral & Diri, 2014; Saad, 2010; Türkmenoğlu, 2015). These steps are briefly explained below.

Tokenization: In this step, the text to be classified is divided into terms by various methods. A term can contain one word or multiple words. Using a method called n-grams, the text can be divided into word-based or character-based terms. In word-based n-gram, the number of n words is treated as a term and in character-based n-gram, the number of n characters is treated as a term.

Normalization: There may be typos in texts on social media. Normalization is a natural language processing procedure that corrects spelling errors. In Turkish text classification studies, normalization is often performed using the Zemberek-NLP natural language processing tool.

Stemming: In the classification of a given text, each word in the text is taken as an attribute; therefore, it is aimed to reduce the number of attributes by stemming. Stemming is a natural language processing operation that is often performed using the Zemberek-NLP natural language processing tool in Turkish (Akın & Akın, 2007).

Stop words removal: While classifying a text, removing stop words is performed to reduce the number of attributes in many studies (Çoban, 2016; Kaynar et al., 2016; Meral & Diri, 2014). Stop words are often those that do not affect the sentiment of the sentence, such as prepositions and conjunctions (Sevindi, 2013).

Term weighting: In the bag-of-word approach, words with high frequency become dominant and cannot provide much information for the model (Waykole & Thakare, 2018). In other words, terms that appear very often in the text may not have any distinctive significance. However, they can have a high weight value. To prevent this, the frequency of the terms is rescaled by the TF-IDF method (TermFrequency – InverseDocumentFrequency) considering how often the terms occur in all texts (Spärck Jones, 2004). The TF value is calculated by the formula in Eq. 1, the IDF value by the formula in Eq. 2, and the TF-IDF by the formula in Eq. 3.

$$tf_{ij} = \frac{F_{ij}}{\sum_i F_j} \tag{1}$$

$$IDF_j = \log(\frac{D}{df_j}) \tag{2}$$

[1] Datashake is a web service that provides reviews and comments from over 85 websites using data scraping. Using the data scraping APIs offered by Datashake, users can easily access reviews and comments on websites like Tripadvisor, Foursquare, and Google Maps (Datashake, 2021).

[2] Data scraping is the process of obtaining desired data from unstructured website content by software using data sets that are suitable for automatic processing (Data Scraping, 2021).

$$w_{i,d} = tf_{ij} * IDF_j \quad (3)$$

where i = text index, j = term index, F = frequency, df_j= number of texts containing j, and D = the number of texts.

Text classification studies in the literature have used different weighting techniques beside TF-IDF such as A-TF, B-TF, LA-TF, L-TF, Knowledge gain, and Chi-square (Sevindi, 2013; Yıldız, 2016).

2.3. Naïve Bayes Classifier

Naïve Bayes is a classification technique based on the probability theory of Bayes (1763). It is based on the assumption that each attribute to be used in classification is independent of each other. It has been used in text classification studies since the early 1960s (Maron, 1961). The Naïve Bayes classifier briefly estimates the class with the highest probability by calculating the probabilities of all cases for each class. The classifier works as follows.

1. Let *X* be a vector that is tried to be predicted and consists of n attributes.

$$X = (x_1, x_2 \dots x_n) \quad (4)$$

2. Let there be m classes in the data set represented by $C_1, C_2 \dots C_m$.
3. The classifier calculates the value with the highest successive probability $P(X|C_i)$ among all classes, as in Eq. 5, to find out which class the vector *X* belongs to.

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)} \quad (5)$$

   a- $P(C_i)$, is calculated as in Eq. 6 by dividing the number of elements in the C_i class by the number of all elements.

$$P(C_i) = \frac{C_i}{|C|} \quad (6)$$

   b- $P(X|C_i)$, is calculated as in Eq. 7 since *X* is an n-element property vector. Since the x_i values are considered independent of each other, there is no need to calculate the *P(X)* value.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (7)$$

4. As a result, the C_i class, which has the largest $P(X|C_i)\,P(C_i)$ value, is determined as the class of *X*.

2.4. Multinominal Naïve Bayes

Multinominal Naïve Bayes is widely used because it is a fast, easy-to-apply, and effective method (Rennie, Shih, Teevan, & Karger, 2003). It is calculated with the formula in Eq. 5 like MNB and NB. In the MNB classifier, different from the NB, the $P(x_k|C_i)$ value is calculated as shown in Eq. 8.

$$P(x_k|C_i) = \frac{N_{ki}}{N_i} \quad (7)$$

where N_{ki} indicates the total frequency of x_k in samples with class C_i containing the x_k attribute. N_k indicates the total frequency of the features included in the samples in the same class (Çoban, 2016; Rennie et al., 2003).

2.5. k-Nearest Neighbor

In this classification method, the k samples nearest to the sample to be classified are calculated by some distance measurement methods and the plural class of the calculated samples is assigned as the class. The k value is a parameter entered by the expert making the classification. Entering a too large k value may cause dissimilar records to be collected together and entering a too small value may cause some records to be assigned to different classes (Khan, Ding, & Perrizo, 2002).

2.6. Support Vector Machines

This is a machine learning algorithm developed by Cortes and Vapnik (1995) for two-group classification problems. Although it was developed for two-group classification problems, it can also be used in multi-group classification problems with planar separation mechanisms in three-dimensional space and hyperplanar separation in multi-dimensional space (Güran, Uysal, & Doğrusöz, 2014).

2.7. Evaluation of Classification Results

In classification processes performed with machine learning, metrics such as accuracy, recall, precision, and f-score are often used to evaluate classification success (Altunkaynak, 2017; Köse, 2018; Parlar & Özel, 2016). These metrics are briefly explained below.

Accuracy: This is a metric that shows what percentage of classified records has been correctly classified. It is obtained by dividing the correct number of classified records by the total number of records. Accuracy is calculated as shown in Eq. 9 in a two-class (positive-negative) classification process (Köse, 2018).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

where TP (True Positive) = number of records correctly classified as positive, TN (True Negative) = number of records correctly classified as negative, FP (False Positive) = number of records falsely classified as positive, and FN (False Negative) = number of records falsely classified as negative.

Recall: Also called the true positive rate (TPR), recall is a metric that shows how many of the true positives were correctly classified. Recall is calculated as shown in Eq. 10 (Köse, 2018).

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

Precision: Precision is calculated as shown in Eq. 11 by dividing the number of correctly classified records by the total estimated number of positives. Precision is a metric that specifies how many of the positives predicted by the classification model were true positives (Shung, 2020).

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

F-Score: This is an evaluation indicator that takes into account both the recall and the precision metrics. It is based on the efficiency criteria of Rijsbergen (1979). It is calculated as shown in Eq. 12 by taking the harmonic means of precision and recall ("F1 score," 2006; Miner et al., 2012).

$$F - Score = \frac{2 * Precision * recall}{Precision + recall} \quad (12)$$

3. Experimental Setup

An experiment was carried out in this section to achieve the aims stated in the introduction. The steps for this experiment are given below.

3.1. Data Preparation

Since the study is a sentiment analysis study, the data should be prepared for analysis. The following operations were carried out for this preparation.

Identifying Turkish Reviews: Since only reviews written in Turkish were to be used in the study, it was necessary to determine the language in which the reviews were written. Many software or services do this. The language values of the data were determined using the language detection method in the Google Translation API[3] and only reviews detected to be in Turkish were used Google, 2019).

Cleaning and Lower Casing: All other characters in the text such as punctuation marks, numbers, and special characters are cleared. All characters are converted to lowercase letters.

Term Normalization: The data is normalized as explained in Section 2.2 using Zemberek-NLP to eliminate spelling errors.

Stemming and Negation: All words are separated using Zemberek-NLP to reduce the number of attributes.

Since Turkish is an agglutinative language, negativity in words is made by suffixes. A negative word can become positive as a result of stemming. Therefore, negative words should be found and appropriate negativity suffixes such as "sız/siz," "me/ma," and "lı/li" should be added to their roots. This process is called negation (Gezici & Yanıkoğlu, 2018). This process was applied to all words that needed to be negated in the sentiment analysis.

Stop Words Removal: Finally, all stop words defined by Sevindi (2013) were deleted. The data were prepared for sentiment analysis. After removing non-Turkish reviews and all stop words, the number of reviews decreased from 49,031 to 37,912.

3.2. Preparation of Data Sets

A number of data sets were created to achieve the aims mentioned in the introduction. To create these data sets, the 37,464 reviews prepared for sentiment analysis in section 3.1 were read one by one by the researcher and the positive, negative, and neutral comments were determined. In other words, manual labeling was performed. The data were also labeled according to the scores. Reviews with scores of four and five were labeled as positive, those with three as neutral, and those with one and two as negative. As a result of the automatic labeling process according to scores, 3458 records were labeled as negative, 5556 as neutral, and 28898 as positive. As a result of manual labeling, 5614 records were labeled as negative, 5658 as neutral, and 26640 as positive.

After the labeling process, data sets were created by taking an equal number of samples from each class. The data sets were created according to the number of classes, the status of labeling, the stems of the words, and negation. The data sets are presented in Table 1. A systematic code was given to each data set for easy use.

As seen in Table 1, data sets are first separated according to their labeling status to determine how the type of labeling affected classification success. Since both two-class and three-class sentiment analysis will be performed with the data sets, they are also divided according to the number of classes. To determine the effect of stemming and negation on classification

[3] Google Translation API is a translation and language detection service that supports over 100 languages (Google, 2019).

Table 1. Data sets used in the study

Labeling Status	Number of Classes	Stemming and Negative Condition	Data Set Code
Labeled Manually	Two (Positive, Negative)	Not Stemmed	M2-DS1
		Stemmed and Not Negated	M2-DS2
		Stemmed and Negative	M2-DS3
	Three (Positive, Negative, Neutral)	Not Stemmed	M3-DS1
		Stemmed and Not Negated	M3-DS2
		Stemmed and Negated	M3-DS3
Auto Labeled by Score	Two (Positive, Negative)	Not Stemmed	S2-DS1
		Stemmed and Not Negated	S2-DS2
		Stemmed and Negated	S2-DS3
	Three (Positive, Negative, Neutral)	Not Stemmed	S3-DS1
		Stemmed and Not Negated	S3-DS2
		Stemmed and Negated	S3-DS3

success, the data sets are divided into three according to the stems of the words and negation.

3.3. Experiment

The data sets were classified using the classic machine learning algorithms NM, MNB, k-NN, and SVM over the WEKA[4] software. The adjustments made in the WEKA software for the classification are listed below.

- The String to Word Vector filter was used to convert the texts to numeric data.
- TF-IDF was used for term weighting.
- The texts in the data sets are divided into word-based n-grams. Tests were conducted with values of 1-gram, 2-gram, and 3-gram.
- In the k-Nearest Neighbor algorithm, the value of k was considered as three, considering the uses in the literature (Silahtaroğlu, 2013). The default parameters of the WEKA software are used in other classification algorithms.
- All data sets were validated by 10-fold cross-validation and classification results were obtained.

4. Results

With the data sets mentioned in the previous section, 144 tests were carried out. With the tests performed, the

Table 2. Classification results of the two-class sentiment analysis process

Labeling Type	Data set	n-gram	NB	MNB	SVM	K-NN
Manually Labeled	M2-DS1	1-gram	0,88	0,93	**0,94**	0,79
		2-gram	0,87	**0,94**	**0,94**	0,63
		3-gram	0,87	**0,94**	0,93	0,62
	M2-DS2	1-gram	0,86	0,92	**0,94**	0,81
		2-gram	0,87	**0,94**	**0,94**	0,71
		3-gram	0,87	**0,94**	**0,94**	0,70
	M2-DS3	1-gram	0,86	0,93	**0,95**	0,81
		2-gram	0,87	0,94	**0,95**	0,70
		3-gram	0,87	**0,94**	**0,94**	0,69
Average F-score			0,87	**0,94**	**0,94**	0,72
Automatically Labeled by Score	S2-DS1	1-gram	0,84	0,87	**0,91**	0,36
		2-gram	0,85	0,87	**0,92**	0,35
		3-gram	0,85	0,87	**0,92**	0,35
	S2-DS2	1-gram	0,86	0,86	**0,91**	0,36
		2-gram	0,86	0,87	**0,93**	0,35
		3-gram	0,86	0,87	**0,93**	0,35
	S2-DS3	1-gram	0,86	0,86	**0,91**	0,36
		2-gram	0,86	0,87	**0,93**	0,35
		3-gram	0,86	0,87	**0,93**	0,35
Average F-score			0,86	0,87	**0,92**	0,35

[4] Weka (Waikato Environment for Knowledge Analysis) is an open-source software used in data mining and machine learning (Witten, Frank, Hall, & Pal, 2016).

Table 3. Contribution of stemming and negation to classification success in two-class sentiment analysis

Data Set	n-gram	NB	MNB	SVM	K-NN
M2-DS2 – M2-DS1 (Stemming)	1-gram	-0,02	-0,01	0	0,03
	2-gram	-0,01	-0,01	0	0,08
	3-gram	-0,01	-0,01	0	0,08
M2-DS3 – M2-DS2 (Negation)	1-gram	0,01	0	0	0
	2-gram	0,01	0	0,01	-0,01
	3-gram	0,01	0	0	-0,01
S2-DS2 – S2-DS1 (Stemming)	1-gram	0,02	-0,01	0	0,01
	2-gram	0,01	-0,01	0	0
	3-gram	0,01	-0,01	0,01	0
S2-DS3 – S2-DS2 (Negation)	1-gram	0	0	0	0
	2-gram	0	0	0	0
	3-gram	0	0	0	0

contribution of labeling, stemming, and negation to classification success were measured, and then the classification success of different classifiers was compared. As a result of these tests, 144 machine learning models were created. The f-score values showing the classification success of machine learning models are given in the tables.

When Table 2 is examined, it is seen that the most successful two-class sentiment analysis classification is performed with the SVM classifier. It is seen that the SVM and MNB algorithms achieved very close classification results with manually labeled data sets. In manually labeled two-class data sets (M2-DS1, M2-DS2, M2-DS3), average f-scores were 0.94 for SVM and MNB, 0.87 for NB, and 0.72 for k-NN. In automatically labeled two-class data sets (S2-DS1, S2-DS2, S2-DS3), average f-scores were 0.92 for SVM, 0.87 for MNB, 0.86 for NB, and 0.35 for k-NN. These results show that classification with manually labeled two-class data sets was more successful than classification with automatically labeled two-class data sets according to scores.

No preprocessing was applied to the data in the data sets ending with DS1. The words in the data sets ending with DS2 were applied stemming. The words in the data sets ending with DS3 were applied both stemming and negation. Besides, each data set was classified into tokens with parameters of one, two, and three grams and the classification process was carried out. The difference between the classification results of data sets ending with DS2 and DS1 shows the contribution of stemming to classification success. Similarly, the difference between the classification results of data sets ending with DS3 and DS2 shows the contribution of negation to classification process. In the light of this information, when Table 3 is examined, it is seen that stemming manually labeled two-class data sets with the k-NN classifier had a significant contribution to classification success. On the other hand, in the classification of automatically labeled two-class data sets with the NB, SVM, and k-NN classifiers, stemming had little contribution. In the classification of manually labeled two-class data sets, negation had a general contribution to classification success. However, the same does not apply to automatically labeled two-class data sets, where negation was found to have no contribution.

In Table 4, the classification results obtained for three-class sentiment analysis are given. When the table

Table 4. Three-class sentiment analysis classification results

Labeling Type	Data set	n-gram	NB	MNB	SVM	K-NN
Manually Labeled	M3-DS1	1-gram	0,70	0,72	**0,79**	0,61
		2-gram	0,70	0,74	**0,80**	0,49
		3-gram	0,70	0,74	**0,79**	0,49
	M3-DS2	1-gram	0,70	0,72	**0,79**	0,64
		2-gram	0,70	0,73	**0,79**	0,55
		3-gram	0,70	0,73	**0,79**	0,54
	M3-DS3	1-gram	0,70	0,72	**0,80**	0,64
		2-gram	0,70	0,74	**0,80**	0,54
		3-gram	0,71	0,73	**0,79**	0,54
Average F-score			0,70	0,73	**0,79**	0,54
Automatically Labeled by Scores	S3-DS1	1-gram	0,59	0,63	**0,70**	0,20
		2-gram	0,59	0,64	**0,70**	0,19
		3-gram	0,59	0,63	**0,69**	0,18
	S3-DS2	1-gram	0,62	0,62	**0,71**	0,22
		2-gram	0,62	0,63	**0,71**	0,19
		3-gram	0,62	0,62	**0,70**	0,19
	S3-DS3	1-gram	0,63	0,63	**0,71**	0,23
		2-gram	0,62	0,63	**0,71**	0,19
		3-gram	0,62	0,63	**0,70**	0,19
Average F-score			0,61	0,63	**0,70**	0,20

Table 2. Contribution of stemming and negation processes to classification success in the three-class sentiment analysis process

Data set	n-gram	NB	MNB	SVM	K-NN
M3-DS2 – M3-DS1 (Stemming)	1-gram	0	-0,01	0	0,03
	2-gram	0	0	0	0,06
	3-gram	0	-0,01	0	0,05
M3-DS3 – M3-DS2 (Negation)	1-gram	0	0	0,01	0
	2-gram	0	0	0,01	-0,01
	3-gram	0	0	0	0
S3-DS2 – S3-DS1 (Stemming)	1-gram	0,03	-0,01	0,01	0,03
	2-gram	0,03	-0,01	0,01	0
	3-gram	0,02	-0,01	0,01	0
S3-DS3 – S3-DS2 (Negation)	1-gram	0,01	0,01	0	0
	2-gram	0,01	0	0	0
	3-gram	0,01	0	0	0

is examined, it is seen that the most successful classifier in the three-class sentiment analysis was the SVM. In manually labeled three-class datasets (M3-DS1, M3-DS2, M3-DS3), average f-scores were 0.79 for SVM, 0.73 for MNB, 0.70 for NB classifier, and 0.54 for k-NN. In automatically labeled three-class data sets (S3-DS1, S3-DS2, S3-DS3), average f-scores were 0.70 for SVM, 0.63 for MNB, 0.61 for NB, and 0.20 for k-NN. These results show that classification with manually labeled three-class data sets was more successful than classification with automatically labeled three-class data sets.

Table 5 shows the contributions of stemming and negation in three-class sentiment analysis to classification success. When Table 5 is examined, it is seen that stemming had a significant contribution to classification success in the classification of manually labeled three-class data sets using the k-NN classifier. However, stemming had little contribution in the classification of automatically labeled three-class data sets using the NB, SVM, and k-NN classifiers. It is also seen that negation contributed to classification success in the processes with the NB, MNB, and SVM classifiers.

5. Conclusion

In this study, a sentiment analysis was carried out with the machine learning method using Turkish reviews for touristic sites. The classification success of the NB, MNB, SVM, and k-NN classifiers was compared in the sentiment analysis, the effects of stemming and negation on classification success were investigated, and the effect of the type of labeling on classification success was measured. Both two-class (positive-negative) and three-class (positive-negative-neutral) sentiment analyses were performed. As a result, the most successful classification result in two-class sentiment analysis was reached using the SVM classifier with an f-score of 0.95. The most successful classification result in three-class sentiment analysis was again reached using the SVM classifier with an f-score of 0.80. It was concluded that the classification results were more successful in the sentiment analysis of manually labeled data sets compared to data sets automatically labeled according to scores. It was determined that stemming significantly contributed to classification success, especially in the k-NN classifier. It was observed that stemming had little contribution to classification success in the NB and SVM classifiers, and a negative effect on classification success in the MNB classifier. Negation resulted in a general increase in classification success in the NB, MNB, and SVM classifiers.

Considering that similar studies in the literature reported f-score values of 0.78 – 0.92 for two-class sentiment analyses and 0.59 – 0.78 for three-class sentiment analyses, it can be said that the results are quite successful (Çoban et al., 2015; Kaya et al., 2012; Kaynar et al., 2016; Velioğlu et al., 2018; Yıldırım et al., 2015). While the findings suggest that the sentiment analysis model created here is feasible for Turkish touristic site reviews, using the data sets in further research and comparing the findings will result in better interpretations.

In the sentiment analysis performed here, the bag-of-word method was used for word representation and classic machine learning algorithms were used as classifiers. Future studies should aim to measure the classification performance of the data sets created here using different word representation methods such as fastText, word2Vec, or glove, along with different machine learning techniques such as artificial neural networks.

References

Akın, A.A., and Akın, M.D., 2018. Zemberek-NLP. Available at: https://github.com/ahmetaa/ zemberek-nlp

Altunkaynak, B., 2017. Veri Madenciliği Yöntemleri ve R Uygulamaları, 2., Seçkin Yayıncılık, Ankara.

Aydoğan, E., and Akcayol, M. A. 2016. A comprehensive survey for sentiment analysis tasks using machine learning techniques. International Symposium on Innovations in Intelligent Systems and Applications, 2-5 August 2016, Sinaia, Romania, pp: 1-7.

Baccianella, S., Esuli, A., and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Seventh

International Conference on Language Resources and Evaluation, 17-23 May, Valletta, Malta, pp: 2200-2204.

Bag-of-Words model 2007. Available at: https://en.wikipedia.org/wiki/ Bag-of-words_model

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Royal Society, 53, 370-418.

Beal, V., n.d. Unstructured data. Available at: https://www.webopedia.com/TERM/U/unstructured_data.html

Bilgin, M., and Şentürk, İ. F. 2017. Sentiment analysis on Twitter data with semi-supervised Doc2Vec. International Conference on Computer Science and Engineering, 5-8 October 2017, Antalya, Turkey, pp: 661-666.

Blumberg, R., and Atre, S., 2003. The problem with unstructured data. Available at: http://soquelgroup.com/wp-content/uploads/2010/01/dmreview_0203_problem.pdf

Can, Ü., and Alataş, B., 2017. Duygu analizi ve fikir madenciliği algoritmalarının incelenmesi. International Journal of Pure and Applied Sciences, 3(1), 75-111.

Chang, Y.-C., Ku, C.-H., and Chen, C.-H., 2019. Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. International Journal of Information Management, 48, 263-279.

Çoban, Ö., 2016. Metin sınıflandırma teknikleri ile türkçe twitter duygu analizi (Master's Thesis), Atatürk University.

Çoban, Ö., Özyer, B., and Özyer, G. T. 2015. Sentiment analysis for Turkish Twitter feeds. 23nd Signal Processing and Communications Applications Conference, 16-19 May 2015, Malatya Turkey, pp: 2388-2391.

Cortes, C., and Vapnik, V., 1995. Support-vector networks. Machine Learning, 20(3): 273-297.

Data Scraping, 2021. Available at: https://en.wikipedia.org/wiki/Data_scraping

Datashake, 2021. Available at: https://www.datashake.com/

Esen, M. F., and Türkay, B., 2017. Turizm endüstrilerinde büyük veri kullanımı. Journal of Tourism and Gastronomy Studies, 4 (4), 92-115.

Esuli, A., and Sebastiani, F., 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. Fifth International Conference on Language Resources and Evaluation, 22-28 May 2006, European Language Resources Association, Geona, Italy, pp: 417-422.

F1 score, 2006. Available at: https://en.wikipedia.org/wiki/F1_score

Gezici, G., and Yanıkoğlu, B., 2018. Sentiment analysis in Turkish. In: Turkish Natural Language Processing, K. Oflazer & M. Saraçlar (Eds.), Springer International Publishing, Cham.

Google, 2019. Google Translation API. Available at: https://cloud.google.com/translate/

Güran, A., Uysal, M., and Doğrusöz, Ö., 2014. Destek vektör makineleri parametre optimizasyonunun duygu analizi üzerindeki etkisi. Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering, 16 (48), 86-93.

Harris, Z. S., 1954. Distributional structure. Word, 10 (2-3): 146-162.

Hearst, M., 2003. What is text mining. Available at: https://people.ischool.berkeley.edu/~hearst/text-mining.html

Kaya, M., Fidan, G., and Toroslu, I. H., 2012. Sentiment analysis of Turkish political news. International Joint Conferences on Web Intelligence and Intelligent Agent Technology, 4-7 December 2012, Macau, China, pp: 174-180.

Kaynar, O., Görmez, Y., Yıldız, M., and Albayrak, A., 2016. Makine Öğrenmesi Yöntemleri ile Duygu Analizi. International Artificial Intelligence and Data Processing Symposium, 17-18 September 2016, Malatya, Turkey, pp: 234-241.

Khan, M., Ding, Q., and Perrizo, W., 2002. K-nearest neighbor classification on spatial data streams using p-trees. 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 6-8 May 2002, Berlin, Germany, pp: 517-518.

Kızılkaya, Y. M., 2018. Duygu analizi ve sosyal medya alanında uygulama (Doctoral Disertation), Bursa Uludağ University.

Köse, İ., 2018. Veri madenciliği teori uygulama ve felsefesi, Papatya Yayıncılık Eğitim, İstanbul.

Ku, L.-W., Liang, Y.-T., and Chen, H.-H., 2006. Opinion extraction, summarization and tracking in news and blog corpora. AAAI Spring Symposium, 27-29 March 2006, CA, USA, pp: 100-107.

Kulcu, S., and Dogdu, E., 2016. A scalable approach for sentiment analysis of Turkish tweets and linking tweets to news. International Conference on Semantic Computing, 4-6 February 2016, CA, USA, pp: 471-476.

Liu, B., 2012. Sentiment analysis and opinion mining. Morgan & Claypool Publishers, Williston.

Liu, K.-L., Li, W.-J., and Guo, M., 2012. Emoticon smoothed language models for twitter sentiment analysis. Twenty-Sixth AAAI Conference on Artificial Intelligence, 22-26 July 2012, Ontario, Canada, pp: 1678-1684.

Lohr, S., 2012. The age of big data. Available at: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

Maron, M. E., 1961. Automatic Indexing: An Experimental Inquiry. J. ACM, 8(3), 404-417.

Meral, M., & Diri, B., 2014. Sentiment analysis on Twitter. 22nd Signal Processing and Communications Applications Conference, 23-25 April 2014, Trabzon, Turkey, pp: 690-693.

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., and Delen, D., 2012. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, Waltham, MA.

Özyurt, B., and Akçayol, M. A., 2018. Fikir madenciliği ve duygu analizi, yaklaşımlar, yöntemler üzerine bir araştırma. Selcuk University Journal of Engineering, Science and Technology, 6(4), 668-693.

Pang, B., Lee, L., and Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. Conference on Empirical Methods in Natural Language Processing, 6-7 July 2002, Philadelphia, USA, pp: 79-86.

Parlar, T., and Özel, S. A., 2016. A new feature selection method for sentiment analysis of Turkish reviews. International Symposium on Innovations in Intelligent Systems and Applications, 2-5 August 2016, Sinaia, Romania, pp: 1-6.

Provost, F., and Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R., 2003. Tackling the poor assumptions of naive bayes text classifiers. International Conference on Machine Learning, 21-24 Agust 2003, Washington, DC, USA, pp: 616-623.

Rijsbergen, C. J. V., 1979. Information Retrieval. Butterworth-Heinemann.

Saad, M. K., 2010. The impact of text preprocessing and term weighting on arabic text classification (Master's Thesis), The Islamic University.

Salur, M. U., Aydın, İ., and Alghrsi, S. A., 2019. SmartSenti: A twitter-based sentiment analysis system for the smart tourism in Turkey. International Artificial Intelligence and Data Processing Symposium, 21-22 September 2019, Malatya, Turkey, pp: 1-5.

Şeker, S. E., 2016. Duygu analizi. Management Information Systems Encyclopedia, 3(3), 21-36.

Sevindi, B. İ., 2013. Comparison of supervised and dictionary based sentiment analysis approaches on Turkish text (Master's Thesis) Gazi University.

Shung, K. P., 2020. Accuracy, Precision, Recall or F1? Available at: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Silahtaroğlu, G., 2013. Veri madenciliği: Kavram ve algoritmaları. Papatya Yayıncılık Eğitim, İstanbul.

Spärck Jones, K., 2004. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 60 (5), 493-502.

Taecharungroj, V., and Mathayomchan, B., 2019. Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. Tourism Management, 75, 550-568.

Toçoğlu, M. A., 2018. Lexicon-based emotion analysis in Turkish (Doctoral Dissertation), Dokuz Eylül University.

Türkmenoğlu, C., 2015. Türkçe metinlerde duygu analizi (Master's Thesis), Istanbul Technical University.

Türkmenoglu, C., and Tantug, A. C., 2014. Sentiment analysis in Turkish media. International Conference on International Conference on Machine Learning, 21-26 June 2014, Beijing, China.

Velioğlu, R., Yıldız, T., and Yıldırım, S., 2018. Sentiment analysis using learning approaches over emojis for Turkish tweets. 3rd International Conference on Computer Science and Application Engineering, 20-23 September, Sanya, China, pp: 303-307.

Waykole, R. N., and Thakare, A., 2018. A Review of feature extraction methods for text classification. IJAERD, 5 (04), 351-254.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J., 2016. The weka workbench. Online appendix for "data mining: practical machine learning tools and techniques". Morgan Kaufmann, Burlington, MA.

Yıldırım, E., Çetin, F. S., Eryiğit, G., and Temel, T., 2015. The impact of NLP on Turkish sentiment analysis. TBV Journal of Computer Science and Engineering, 7(1), 41-51.

Yıldız, O., 2016. Metin madenciliğinde anahtar kelime seçimi bir üniversite örneği, Journal of Management Information Systems, 2(1), 29-50.