

## Research Article

Özge Barış-Tüzemen, Samet Tüzemen, Ali Kemal Çelik\*

# Sentiment analysis of reviews on cappadocia: The land of beautiful horses in the eyes of tourists

<https://doi.org/10.2478/ejthr-2023-0015>

received September 6, 2023; accepted August 28, 2023

**Abstract:** The Cappadocia region is one of the most popular tourist destinations in Turkey, and its tourism sector has a significant share in the Turkish economy. In this study, we scraped TripAdvisor reviews of visitors of the Cappadocia region with the Python programming language and used them to analyse public sentiment using various supervised machine learning algorithms. The main purpose of the study is to help create competitive intelligence on both regional and global scales using social media data. For this, we applied Random Forest, Naïve Bayes, and Support Vector Machine methods to classify 4,770 reviews and get insights about the visitors' perspectives. Results show that the majority of the tourists (90%) had a positive experience during their visit. Most of the complaints focused on the attitudes of staff members. In addition, all three supervised machine learning methods achieved high accuracy in their classification of the reviews. This study is significant in terms of providing a meaningful database for understanding visitor comments, the most important data for the development of tourism in the region, through state-of-the-art machine learning methods, and to direct improvements accordingly.

**Keywords:** Sentiment Analysis; Supervised Machine Learning; Tourism Industry; Cappadocia; Random Forest; Naïve Bayes

## 1 Introduction

Social media has an important role as a source of information for travellers around the world. Social networks have led to widespread changes in “business-to-business”, “business-to-customer” and “customer-to-customer” communications. The Internet has evolved from a broadcast medium to a participatory platform that allows people to themselves be the “media” for collaboration and information sharing (Leung et al., 2013). According to a 2021 Statista report, there were 859 million reviews on TripAdvisor as of 2019 (Statista, 2021b). In July 2019 alone, 224 million people visited the TripAdvisor website. The website Booking.com, another popular and well-known travel platform, was visited 697 million times in the same period (Condor Ferries, 2021). In addition, Twitter had an average of 330 million active users per month in 2019 (Statista, 2021a).

These statistics clearly show that the dominance of social media in our lives has increased. With this increase, it is seen that corporate communication has become democratized (Kietzmann et al., 2011). The ability to share information through social media has created significant changes in the bargaining power of consumers and decreased information asymmetries. Stories and experiences shared by people across many online platforms (blogs, microblogs, Twitter, Facebook, TripAdvisor, Booking.com, YouTube, etc.) have made social media into a megatrend, significantly affecting the tourism sector (Leung et al., 2013).

The tourism industry, whose stakeholders are more than in many other industries, is the largest job-creating sector on the planet. Developing technology and the internet play a significant role for tourism organizations and destinations in increasing their market share. Advances in search engines and social networks have increased the number of people who plan and experience their travels around the world. These developments have also affected the efficiency of tourism organizations and their business models, as well as the way consumers communicate with these organizations. There have been many new players

\*Corresponding author: Ali Kemal Çelik, Ardahan University, Department of Quantitative Methods, Ardahan, Turkey, Email: alikemalcelik@ardahan.edu.tr

Özge Barış-Tüzemen, Uppsala University, Department of Statistics, Uppsala, Sweden

Samet Tüzemen, Ardahan University, Department of Business Administration, Turkey

in the tourism sector since these developments, and the importance of tourism for a growing number of national and regional economies has been increasingly recognized (UNWTO, 2001; Buhalis & Law, 2008). In short, the internet and social networks have significantly changed the distribution channels of tourism-related information, as well as travel planning and consumption habits (Xiang & Gretzel, 2010).

Tourism is both a service and a knowledge-intensive industry. Therefore, it is essential to understand how changes in technologies and consumer behaviour are affecting the distribution and availability of travel-related information. Technology and social media are needed to create an avant-garde tourism experience. Information and communication technologies are powerful tools for reshaping tourism as they create totally new products, communication networks, business models, industry structures, and types of companies. Therefore, they are an important resource for travel companies seeking an enhanced competitive advantage and maximum profit in the global market (Sheldon, 2006).

There is extensive literature on social media analytics that combines web scraping, machine learning applications with recent software, and statistical techniques to collect, clean, analyse, and understand large amounts of data. With these methods, opinions and beliefs about products can be examined for commercial purposes, such as tracking and identifying trending topics and popular sentiment (Xiang et al., 2017). In particular, with social media playing an increasingly in both travellers' decisions and tourism activities and management, studies are increasingly focusing on the effects of user ratings and comments on tourism destinations and hotel management practices (see Vermeulen and Seegers, 2009; Ye et al., 2009; Xiang et al., 2008; Schmunk et al., 2014; Schuckert et al., 2015; Fang et al., 2016; Geetha et al., 2017; Cheng and Jin, 2019; Kyaw and Wai, 2019; Nohh et al., 2019; Büyükeke et al., 2020; Dina, 2020; Paolanti et al., 2021; Özen, 2021).

Turkey is in the Mediterranean climate zone, and its clean beaches and coves, historical sites, and natural beauty all endow it with great potential for tourism. Turkey has hosted many civilizations throughout history and offers a wide range of options to visitors of all preferences and tastes. Each one of its seven regions has its own unique natural sites. Therefore, the development of the tourism sector could play an important role in solving socio-economic problems in Turkey, such as unemployment, inflation, and migration (Bulut, 2018). Turkey, which welcomed 51.2 million tourists in 2019, is ranked 6th in the world in terms of number of arrivals. However, the country ranks 13th in the world in terms of tourism reve-

nues, which stood at 29.8 billion dollars in 2019 (Republic of Turkey Ministry of Culture and Tourism, 2020).

The Cappadocia region, which encompasses the Nevsehir, Aksaray, Nigde, Kayseri, and Kirsehir provinces, is one of Turkey's most important tourism destinations. The Rocky Cappadocia region, which is a narrower area, consists of Uchisar, Urgup, Avanos, Goreme, Derinkuyu, Kaymakli, Ihlara, and their surroundings. It is an important tourism destination of both natural and historical interest and has been listed as a UNESCO World Heritage Site since 1985. Natural and cultural tourism is popular in the region, encompassing experiences such as thermal baths, ballooning, horse-back riding, and pottery, just to name a few (Ahiler Development Agency Plans for Future, 2015; Republic of Turkey Ministry of Culture and Tourism, 2021). Common activities include hiking and horseback riding in Ihlara; visiting the Red and Rose Valleys, or the Göreme or Zelve open-air museums; and taking a hot air balloon tour over the region.

Analysing social media data in the tourism industry will allow both local businesses and policymakers to understand the decision-making behaviours of visitors and to be aware of the opportunities and pitfalls in the industry. In this context, we investigated the social media data of the Rocky Cappadocia region through sentiment analysis and various machine-learning methods.

In the next part of the study, we briefly introduce social media and sentiment analysis methods. Later, we explain sentiment classification and machine learning algorithms. In the fourth section, we discuss the data set and findings, and finally, we present the conclusion and evaluations.

## 2 Social Media and Sentiment Analysis

When using social media, people interact with each other and share information, opinions, and many other things freely. At the same time, companies and organizations may reach customers more easily and enjoy advantages such as offering them personalised options. According to Appel et al. (2019), social media sites should be seen as "digital places" where people manage important parts of their lives, rather than just platforms that offer digital media and technology services. When evaluated from this perspective, social media will mean "less about specific technologies or platforms and more about what people do in these environments" (Appel et al., 2019, p. 80).

On the other hand, reviews made by social media users cannot be treated as the single source of truth. False information can spread very quickly on social media. Thus, users need to pay attention to this kind of information and always double-check the information they find on social media.

Sentiment analysis is a method for analysing emotions by evaluating posts on internet platforms. Making inferences from the related platforms on the feelings and opinions of people provides an understanding of human behaviour, with implications for efficiency in areas such as international public influence, business decisions, and policy development (Alamoodi *et al.*, 2020). Sentiment analysis has many other applications in education, health, and tourism management, among many other areas (Balahadia *et al.*, 2016; Alayba *et al.*, 2017; Ye *et al.*, 2009; Saura *et al.*, 2019). It is fast and effective for making large-scale inferences and decisions at a low cost (Choi *et al.*, 2017). For these reasons, it is an essential tool for understanding social trends.

### 3 Methodology

To analyse the general public opinion about the Cappadocia region, we used sentiment analysis, an effective and frequently-used method in big-data analytics. In sentiment analysis, the main idea in a text is classified by applying natural-language processing and text analytics. Sentiment analysis aims to understand the attitude of the author by detecting the emotional polarity of a text and classifying it as positive, negative, or neutral (Luo *et al.*, 2013, pp. 53–54). For this, a dictionary-based emotion score is determined for each word in the text using flexible and open-source programming languages such as R, Python, and related packages. Later, this score, determined on the basis of the selected words, is calculated for the whole text. As a result, the entire text can be classified as positive, negative, or neutral (Luo *et al.*, 2013).

Sentiment analysis studies are generally performed with three different methods: the lexicon-based method, the machine learning method and the hybrid approach. The lexicon-based approach “detects the sentiment based on a sentiment lexicon, which includes a collection of known and pre-compiled sentiment terms. It is divided into two, namely dictionary-based approach and corpus-based approaches, using statistical or semantic methods to find sentiment polarity” (Medhat *et al.*, 2014, p. 1098). Machine learning approaches apply machine learning algorithms by taking advantage of syntactic

and/or linguistic features, including sentiment lexicons (Maynard & Funk, 2011). The hybrid approach combines both machine learning and a lexicon-based method with manually written linguistic rules. The classifiers in this approach are used gradually. Thus, if one method cannot classify a document, the algorithm passes to the next method until the document is classified or all methods are tried (Prabowo & Thelwall, 2009; Manda, 2019). In this study, opinions of tourists about the Cappadocia region, which is one of the most popular tourist destinations in Turkey, is examined using machine learning algorithms.

### 3.1 Machine Learning

Machine learning techniques for sentiment classification have gained attention because of the possibility of modelling many features while also capturing the context, adapting easily to changing inputs, and measuring the degree of uncertainty with which a classification is made (Boiy & Moens, 2008). There are two basic kinds of machine learning algorithms: supervised and unsupervised learning. “A supervised learning algorithm takes a known set of input data (training set) and known responses to data (output) and trains a model to generate reasonable predictions for response to new input data.” (MathWorks, 2016, p.2). Unsupervised learning involves pattern recognition without the involvement of a target feature. In other words, all variables are used as features (Alloghani *et al.*, 2019). Three different supervised machine learning algorithms are employed in the study. These classification algorithms are Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF).

#### 3.1.1 Naïve Bayes (NB)

Naïve Bayes, which is the simplest example of a probabilistic classifier, is the probability ( $P(C | d)$ ) that a  $d$  document belongs to a  $C$  class. The Naïve Bayes Classifier does not take the possible dependencies between inputs into account and reduces a multivariate problem into a set of univariate problems. With supervised training, Naïve Bayes learns the pattern of reviewing a set of well-categorized test documents, thus comparing the contents of all categories with a word list and their emergence. Such word formation lists are used to classify new documents into the correct sub-categories according to the highest-end probability (Islam *et al.*, 2007; Ting *et al.*, 2011).

### 3.1.2 Support Vector Machine (SVM)

“Support Vector Machine is a supervised machine algorithm that uses statistical learning theories. It works as separating the classes in the dataset with an optimal hyperplane that maximizes the margin between classes” (Yu et al., 2012, p.232). In other words, SVM maximizes the distance between support vectors belonging to different classes. “The interesting property of SVM is that it is an approximate implementation to the structure risk minimization (SRM) in statistical learning theory, rather than the empirical risk minimization method from which the classification function is derived by minimizing mean square error (MSE)” (Song et al., 2002, p.440). As a powerful machine learning algorithm, SVM works well in text classification, as it has good generalization ability in a high-dimensional feature space (Joachims, 1998; Wei et al., 2012; Nohh et al., 2019).

### 3.1.3 Random Forest (RF)

Random forests are an important bagging modification that create a collection of de-correlated trees and then averages them. Random forests are used to improve bagging variance reduction by scaling down the correlation among trees without changing the variance too much. This is done by randomly selecting input variables during the tree growing process. “On many problems, the performance of random forests is very similar to boosting, and they are simpler to train and tune” (Hastie et al., 2009, pp. 587–588). “It gives good results in data sets that contain categorical variables with a large number of variables and class labels, have missing data or exhibit an uneven distribution” (Aydm, 2018, p.172).

## 3.2 Performance Evaluation

In the current study, the success of classification algorithms is decided by confusion matrix, mean absolute

error (MAE) and root mean square error (RMSE) criteria. The confusion matrix gives the actual and predicted classification data computed from a classification system. The performance of these kinds of systems is often determined using the data in the matrix. A two-class classification problem confusion matrix is shown below:

Accuracy is the statistic that determines the success of the machine learning model. As seen in Equation (1), the accuracy rate is the ratio of data classified as correct to all data:

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

Precision, Recall, and F1-score are three other criteria that can be used to evaluate the outcome of emotion analysis. Considering an exemplary system of sentiment analysis that has two classes and using the notations given in Table 1, Precision is calculated as seen in Equation (2). Precision statistics of a positive class are calculated by dividing correctly classified samples to the total samples predicted to be class-positive. The Recall, on the other hand, is calculated by dividing correctly classified positives (True Positives) by the sum of correctly classified positives (True Positives) and incorrectly classified negatives (False Negatives), as seen in Equation (3). Finally, the F1-score is the weighted harmonic average of Precision and Recall, as seen in Equation (4). According to these definitions as they are explained, Precision, Recall and F1 values of the positive class are calculated as follows (Ribeiro et al., 2016; Alaei et al., 2017):

$$Precision_{(Positive)} = \frac{TP}{TP+FP} \tag{2}$$

$$Recall_{(Positive)} = \frac{TP}{TP+FN} \tag{3}$$

$$F1_{(Positive)} = \frac{2xPrecision_{(Positive)}xRecall_{(Positive)}}{Precision_{(Positive)}+Recall_{(Positive)}} \tag{4}$$

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) statistics are also used in this study to evaluate the performance of classification algorithms. “The mean

**Table 1:** The Confusion Matrix for Two-Class Classification Problem

Predicted	Actual	
	“Positive”	“Negative”
“Positive”	TP	FP
“Negative”	FN	TN

“Note: TP is True Positive, FP is False Positive, FN is False Negative, and TN is True Negative.” Source: Visa et al. (2011)

absolute error expresses the mean of the verification sample of the absolute values of the differences between the estimated and observed,” as is seen in Equation (5) below (Yi & Liu, 2020, p.632):

$$MAE = \frac{1}{n} \sum_{i=1}^n |d_i - \hat{d}_i| \quad (5)$$

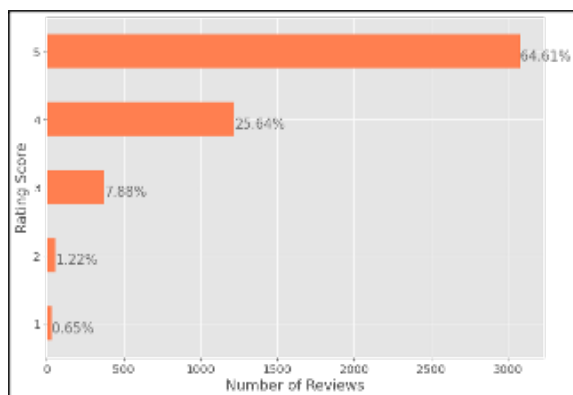
where  $d$  is the observed and  $\hat{d}$  is the estimated value. A smaller MAE value indicates better prediction accuracy. RMSE is a quadratic scoring rule that measures the mean error size. Each difference between the estimated and the observed values is squared and then averaged on the sample, as seen in Equation (6) below (Aydın, 2018):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (d_i - \hat{d}_i)^2} \quad (6)$$

RMSE should be equal to or greater than MAE. The magnitude of the variance between the individual errors in the sample will be parallel to the size of the gap between RMSE and MAE. On the other hand, if MAE is equal to RMSE, it means that all the errors have the same size.

## 4 Data Set

In this study, we collected social media reviews on Göreme National Park and Göreme Open Air Museum in the Rocky Cappadocia region from the TripAdvisor website. The collection of reviews, also known as data scraping, was done using the Python programming language. A total number of 4,770 English-language reviews were collected from April 2011 to January 2021. In order to analyse the reviews, first we converted the text into numbers. To do that, we used the “term frequency-inverse document frequency” (TF-IDF) method, which supposes that a document is just a stack of words. Therefore, the relative importance of any word in the document can be calculated, that is, vec-



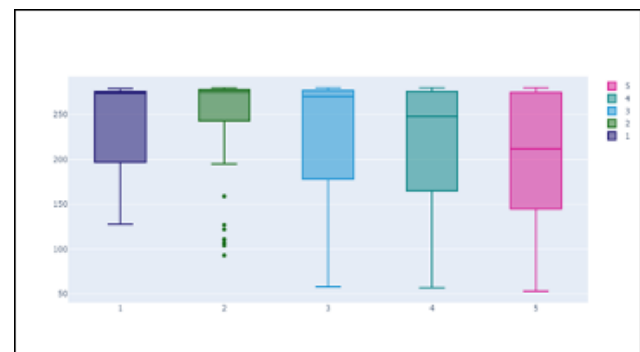
**Figure 1:** Distribution of Reviews According to Rating  
Source: Authors' computation.

torized, by considering the frequency of the word in the document and its popularity in the compilation (Kim et al., 2019, p.17). We retained only the most common 2,000 words in the dataset, as less common words do not play a crucial role in classification. In addition, a word needed to be in at least five reviews, and at most in 70% of all the reviews. The rationale behind choosing 70% as a threshold was that words that occur in more than 70% of documents are very common and unlikely to play any role in classifying sentiment. Finally, we created a term document matrix using the CountVectorizer module.

We applied data cleaning before creating the term document matrix. After that, we converted all characters in the reviews to lowercase and removed punctuation and numeric characters. Then, we removed the words that do not significantly affect the meaning of the sentence using the StopWords application and lemmatized the text to map the various forms of the words to the root form using WordNet Lemmatizer. Two classifications, positive and negative, were made according to the ratings of the reviews. We labelled reviews that have ratings of 1, 2 and 3 as negative, and 4 and 5 as positive. 70% of the data set was used as the training set, and the remaining part was used for the test set.

## 5 Findings

In this part of the study, we make conclusions about the machine learning algorithm results and interpretations regarding the sentiment polarity of tourists who visited Göreme Open Air Museum and Göreme National Park. In this context, first, the distribution of reviews by rating is shown in Figure 1. According to Figure 1, the tourist reviews of Cappadocia generally show satisfaction with the experience that they have had. In the rating distribution of the reviews, the comments with ratings of 4 and 5



**Figure 2:** Review Length by Rating  
Source: Authors' computation.

constitute approximately 90% of the total reviews, while the remaining ratings compose 10%.

Figure 2 shows the length of the reviews according to the rating. 5-star reviews have the lowest median size, while the highest median sizes are for 1 and 2-star reviews. In addition, considering the minimum number of characters, it can be observed that 1- and 2-star comments are longer than those given other rating values. The greater length of the negative reviews is included in the figure.

In Table 2, some of the negative and positive reviews of the Cappadocia Region are given. As can be seen from the table, positive reviews are generally shorter than negative ones. On the other hand, staff is the aspect of the location most frequently complained about in negative reviews. This indicates that the businesses in the region should take steps to increase the quality of their service by taking into account these reviews.

The most frequently mentioned bigrams all appear in the reviews in Figure 3. In fact, the most common bigrams in the text are “open air,” “national park,” and “goreme air.” However, since the reviews of Göreme Open Air Museum and Göreme National Park are being exam-

ined within the scope of the study, these words, which are likely to be seen in every review, have been removed from the chart. Apart from this, one of the most important buildings of Göreme Open Air Museum, the Dark Church, is the most mentioned bigram in the reviews. Rocks composed of lava and ash are also frequently mentioned in the reviews.

Finally, Figure 4 shows the frequency of the 20 most common trigrams in the reviews. According to the chart, tourists like the Cappadocia Region and it is seen that the Dark Church is mentioned a lot in trigrams as well. The popularity of the hot air balloon ride, one of the important activities in the region, is also reflected in the reviews, and the local people living in caves were also the focus of tourists’ interest.

Performance evaluation of the classification algorithms is examined using the micro-average values of F1-Score, Recall, and Precision statistics, and the outcome is presented in Table 3. While the macro-average treats all classes equally, the micro-average supports larger classes. Thus, in cases where there are class imbalances, the

**Table 2:** Reviews and Rating

Heading	Review	Rating
“DISSAPOINTED WITH THE STAFF – VERY RUDE, RUDE STAFF.”	“My son had to use the toilet as soon as we entered. So we went back out to the toilets that were at the entrance in the shop. They wouldnt let us back in, they wanted us to buy new tickets. They even called the ploice on us.”	1
“A total scam”	“We paid £200 expecting a private driver with a English speaking guide instead we got a miserable driver that didn’t even communicate with us and didn’t even have the curtesy to even get out of the car...no guide and we were left to our own devices!!!completely ruined our whole...”	1
“Wonderfull but local authorities dont take care of this museum,”	“I was shocked when I saw such a piece of neglected history and no care from yhe local authorities that made me almost call the police. People taking peaces of the caves and taking home can you beleive it?Apart of it certainly it a place to visit.”	1
“Disappointed with the visit”	“We visited Zelve Open Air Museum the day before, which is why the Goreme Open Air Museum was very disappointing for us. By comparison, GoremeOAM was 3x more expensive than Zelve, most of the sites inside Goreme OAM was closed, too many people were around...”	1
“Beautiful,Very beautiful place.”	“I went there 2 months ago. It was great trip. My guide told me history about the place. It is really great”	5
“Amazing”	“One of the best place in Cappadocia, let me write correctly, It is one of the best place in World.”	5
“Stunning,”	“Absolute stunning frescoes in the dark church... Well worth the extra 8 lira, A must see in Cappadocia and an easy walk from the towns center”	5
“Amazing,”	“Göreme Open Air Museum is a magnificent open air park, ideal for a guided tour. Many caves are great for exploring. It is good to see ‘churches’ in some of these caves.”	5

Source: TripAdvisor

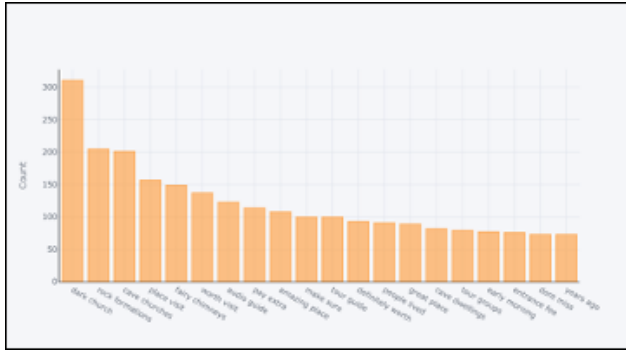


Figure 3: Top 20 Bigrams in Reviews

Source: Authors' computation.

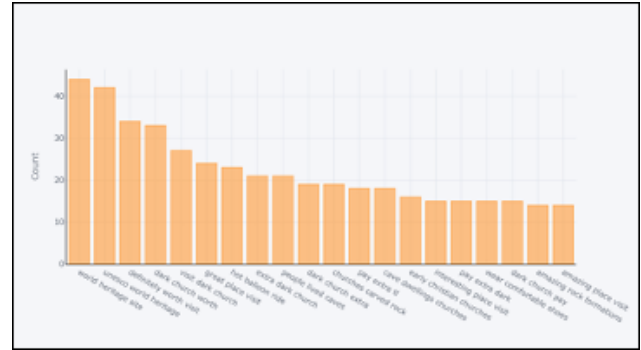


Figure 4: Top 20 Trigrams in Reviews

Source: Authors' computation.

micro-average value is preferred (Sokolova & Lapalme, 2009; Nohh et al., 2019). sssss

Table 3 also shows the RMSE and MAE results. The degree of success achieved by each of the three algorithms is very similar. However, Naïve Bayes algorithm performed the classification process with the highest accuracy rate. At the same time, the lowest RMSE and MAE values belonged to the Naïve Bayes algorithm. Based on these results, we found the Naïve Bayes algorithm to be preferable for classifying the reviews.

## 6 Discussion

In this study, we analysed sentiment about the Cappadocia region, which is one of Turkey's most important tourism destinations. We employed various machine learning algorithms and visualization methods for this purpose. We collected 4,770 English reviews for Göreme Open Air Museum and Göreme National Park in the Cappadocia region. For classification purposes, we carried out Random Forest, Naïve Bayes, and Support Vector Machine supervised machine learning methods. We used accuracy, RMSE and MAE criteria for performance evaluation. Although the classification algorithms performed very similarly, we determined that the Naïve Bayes algorithm was the most successful among them.

Most of the reviews (90%) indicated that their writers had positive feelings about the region. Tourism revenues in both the regional and the national economy are one of the most important components of Turkey's GDP. Cappadocia, with its fairy chimneys, valleys, underground cities, churches, inns, and cultural structures, carries the traces of a civilization thousands of years old. It can be seen from the chart in the findings section, churches are mentioned in most of the reviews, meaning religious tourism is popular. The Dark Church, in particular, was at the centre of tourists' attention. Strategic plans should be made to attract more visitors from various parts of the world by preserving the spirit and potential of the Cappadocia Region. The number and focus of negative reviews show that it is easy to eliminate the negative factors with some small improvements in customer service. The success of the tourism sector will eventually contribute to the income level of the locals as well.

## 7 Conclusion

Language is a powerful mechanism that enables people to express themselves. Studies on sentiment analysis, also known as opinion mining, are one of the most interesting and rapidly-growing research areas of natural language processing and text mining. A wide range of sentiment

Table 3: Performance Evaluation

Classifier	Accuracy (%)	"Precision"	"Recall"	"F1-Score"	"RMSE"	"MAE"
"RF"	90.91	0.9091	0.9091	0.9091	0.7764	0.18169
"SVM"	90.98	0.9098	0.9098	0.9098	0.7749	0.1802
"NB"	91.05	0.9105	0.9105	0.9105	0.7734	0.1788

Source: Authors' estimation.

analyses can be made for every field through social media reviews, which are a crucial artifact of the information age. Social media allows individuals and organizations to reach larger audiences at much less cost than surveys, interviews, and similar methods. In addition, on social platforms, people share their feelings and thoughts in line with their own wishes. This causes these opinions to be more candid than others gathered by more conventional methods. It is thought that sentiment analysis on the tourism sector using social media reviews will help countries and businesses make better strategic decisions for the future. In addition, it will prevent the information-asymmetry problem in the global competitiveness, and thus enhance awareness of opportunities and threats in the market.

This study reveals the overall opinion of the visitors about Cappadocia region and thus serves as a mirror for service providers in the region. To be able to improve the quality perceptions of the region, locals, service providers and decision makers should take these reviews into account and keep increasing service quality accordingly. A significant limitation of this study, however, is that only a certain part of the Cappadocia Region is investigated. Thus, as a suggestion for further investigations, expanding the research to the whole region, or to other tourist destinations, may further reveal and contribute to the tourism potential of Turkey.

## Conflicts of Interest

The authors of the article “Sentiment analysis of reviews on Cappadocia: the land of beautiful horses in the eyes of tourists” declare no conflict of interest.

## Bionotes

**Özge Barış Tüzemen**, PhD, is a visiting researcher at Uppsala University, Sweden. Her main research areas are econometrics and statistics. She has published in many top-tier journals, including *Environmental Science and Pollution Research and Sustainable Cities and Society*. Orcid ID: 0000-0001-7692-5032

**Samet Tüzemen**, PhD, is a research assistant at Ardahan University, Turkey. His main research area is econometrics. He has published in many top-tier journals, includ-

ing *Environmental Science and Pollution Research*. Orcid ID: 0000-0003-1465-4489

**Ali Kemal Çelik**, PhD, is an Associate Professor of Quantitative Decision Methods at Ardahan University, Turkey. His main research interests are discrete-choice modeling and transportation behaviour. He has published in many top-tier journals, including *Accident Analysis & Prevention, Energy & Buildings, Environmental Science and Pollution Research, Transportation Letters, International Journal of Occupational Safety and Ergonomics, Ekonomiska Istraživanja, Transylvanian Review of Administrative Sciences, Transportation Research Record: Journal of the Transportation Research Board, Romanian Journal of Economic Forecasting, Borsa Istanbul Review, Estudios de Economía, and European Research on Management and Business Economics*. Orcid ID: 0000-0002-2605-6526

## References

- [1] Ahiler Development Agency Plans for Future. (2015). *Tourism and Investment Opportunities in Cappadocia*. <https://www.ahika.gov.tr/>
- [2] Alaei, A. R., Becken, S., & Stantic, B. (2017). Sentiment analysis in tourism: Capitalizing on Big Data. *Journal of Travel Research*, 58(2), 1–17. <https://doi.org/10.1177/0047287517747753>
- [3] Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., Hameed, H., & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, 167(2021), 1s13. <https://doi.org/10.1016/j.eswa.2020.114155>
- [4] Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017, April 3-5). *Arabic Language Sentiment Analysis on Health Services* [Conference Session]. International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France. <https://ieeexplore.ieee.org/document/8067771#full-text-section>
- [5] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2019). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In M. W. Berry, A. Mohamed & B. W. Yap (Eds.), *Supervised and Unsupervised Learning for Data Science: Unsupervised and Semi-Supervised Learning* (pp.3–22). Springer. [https://doi.org/10.1007/978-3-030-22475-2\\_1](https://doi.org/10.1007/978-3-030-22475-2_1)
- [6] Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing Science*, 48(2020), 79–95. <https://doi.org/10.1007/s11747-019-00695-1>
- [7] Aydın, C. (2019). Classification of the fire station requirement with using machine learning algorithms. *International Journal of Information Technology and Computer Science*, 11(1),

- 24–30. <https://www.mecs-press.org/ijitcs/ijitcs-v11-n1/IJITCS-V11-N1-3.pdf>
- [8] Balahadia, F. F., Fernando, G. G., & Juanatas, I. C. (2016, May 9-11). *Teacher's performance evaluation tool using opinion mining with sentiment analysis* [Conference Session]. IEEE Region 10 Symposium (TENSYP), Bali, Indonesia. <https://ieeexplore.ieee.org/document/7519384>
- [9] Boiy, E., & Moens, M.-F. (2008). A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, 12(2008), 526–558. <https://doi.org/10.1007/s10791-008-9070-z>
- [10] Buhalis, D., & Law, R. (2008). Progress in Information Technology and Tourism Management: 20 Years on and 10 Years After the Internet – The State of eTourism Research. *Tourism Management*, 29(4), 609-23. <https://doi.org/10.1016/j.tourman.2008.01.005>
- [11] Bulut, R. (2018). Tourism Sector Economic Contribution. *Ayrıntı: Göller Yöresi Ekonomi ve Kültür Dergisi (Detail: Lakes Region Economy and Culture Magazine)*, 5(62), 54–57.
- [12] Büyükeke, A., Sökmen, A., & Gencer, C. (2020). Metin madenciliği ve duygu analizi yöntemleri ile sosyal medya verilerinden rekabetçi avantaj elde etme: Turizm sektöründe bir araştırma (Gaining competitive advantage from social media data with text mining and sentiment analysis methods: A research in the tourism sector). *Journal of Tourism and Gastronomy Studies*, 8(1), 322–335. [https://www.researchgate.net/publication/340324271\\_Metin\\_Madenciligi\\_ve\\_Duygu\\_Analizi\\_Yontemleri\\_ile\\_Sosyal\\_Medya\\_Verilerinden\\_Rekabetci\\_Avantaj\\_Elde\\_Etme\\_Turizm\\_Sektorunde\\_Bir\\_Arastirma\\_Gaining\\_Competitive\\_Advantage\\_from\\_Social\\_Media\\_Data\\_with\\_Text\\_M](https://www.researchgate.net/publication/340324271_Metin_Madenciligi_ve_Duygu_Analizi_Yontemleri_ile_Sosyal_Medya_Verilerinden_Rekabetci_Avantaj_Elde_Etme_Turizm_Sektorunde_Bir_Arastirma_Gaining_Competitive_Advantage_from_Social_Media_Data_with_Text_M)
- [13] Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, 76(Part A), 58–70. <https://doi.org/10.1016/j.ijhm.2018.04.004>
- [14] Choi, S., Lee, J., Kang, M.-G., Min, H., Chang, Y.-S., & Yoon, S. (2017). Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods*, 129(2017), 50–59. <https://doi.org/10.1016/j.ymeth.2017.07.027>
- [15] Condor Ferries. (2021). *Online Travel Booking Statistics - Research & Planning*. <https://www.condorferries.co.uk/online-travel-booking-statistics>
- [16] Dina, N. Z. (2020). Tourist sentiment analysis on TripAdvisor using text mining: A case study using hotels in Ubud, Bali. *African Journal of Hospitality, Tourism and Leisure*, 9(2), 1–10. [https://www.ajhtl.com/uploads/7/1/6/3/7163688/article\\_12\\_vol\\_9\\_2\\_2020\\_indonesia.pdf](https://www.ajhtl.com/uploads/7/1/6/3/7163688/article_12_vol_9_2_2020_indonesia.pdf)
- [17] Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*, 52(2016), 498-506. <https://doi.org/10.1016/j.tourman.2015.07.018>
- [18] Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels – An empirical analysis. *Tourism Management*, 61(2017), 43-54. <https://doi.org/10.1016/j.tourman.2016.12.022>
- [19] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. <https://link.springer.com/book/10.1007/978-0-387-84858-7>
- [20] Islam, M. J., Wu, Q. M., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November 21-23). *Investigating the performance of Naïve-Bayes classifiers and K-Nearest Neighbor classifiers* [Conference Session]. International Conference on Convergence Information Technology. <https://ieeexplore.ieee.org/xpl/conhome/4420216/proceeding>
- [21] Joachims, T. (1998, April 21-23). *Text categorization with support vector machines: learning with many relevant features* [Conference Session]. European Conference on Machine Learning, Chemnitz, Germany. <https://link.springer.com/chapter/10.1007/BFb0026683>
- [22] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social Media? Get Serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241-51. <https://doi.org/10.1016/j.bushor.2011.01.005>
- [23] Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Information Sciences*, 477(2019), 15–29. <https://doi.org/10.1016/j.ins.2018.10.006>
- [24] Kyaw, N., & Wai, T. (2019, November 6-7). *inferring user preferences using reviews for rating prediction* [Conference Session]. International Conference on Advanced Information Technologies (ICAIT), Yangon, Myanmar. <https://ieeexplore.ieee.org/document/8921179>
- [25] Leung, D., Law, R., Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: a literature review. *Journal of Travel & Tourism Marketing*, 30(1-2), 3–22. <https://doi.org/10.1080/10548408.2013.750919>
- [26] Luo, T., Chen, S., Xu, G., & Zhou, J. (2013). Sentiment analysis. In T. Luo, S. Chen, G. Xu & J. Zhou (Eds.), *Trust-based Collective View Prediction* (pp. 53–68). Springer. [https://doi.org/10.1007/978-1-4614-7202-5\\_4](https://doi.org/10.1007/978-1-4614-7202-5_4)
- [27] Manda, K. R. (2019). *Sentiment Analysis of Twitter Data Using Machine Learning and Deep Learning Methods*. [Master's thesis, Blekinge Institute of Technology.] DiVA portal. <https://www.diva-portal.org/smash/get/diva2:1335995/FULLTEXT02>
- [28] MathWorks. (2016). *Applying Supervised Learning*. [https://www.mathworks.com/content/dam/mathworks/tag-team/Objects/i/90221\\_80827v00\\_machine\\_learning\\_section4\\_ebook\\_v03.pdf](https://www.mathworks.com/content/dam/mathworks/tag-team/Objects/i/90221_80827v00_machine_learning_section4_ebook_v03.pdf)
- [29] Maynard, D., & Funk, A. (2011, May 29 – June 2). *Automatic detection of political opinions in Tweets* [Conference Session]. 8<sup>th</sup> Extended Semantic Web Conference, Heraklion, Greece. [https://doi.org/10.1007/978-3-642-25953-1\\_8](https://doi.org/10.1007/978-3-642-25953-1_8)
- [30] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- [31] Nohh, N. H., Zainuddin, N. M., Anuar, S., Azmi, N. F., & Hassan, W. A. (2019). Sentiment analysis towards hotel reviews. *sOpen International Journal of Informatics*, 7(Special Issue 2), 1–19. <https://oiji.utm.my/index.php/oiji/article/view/74>
- [32] Özen, A. (2021). Evaluation of tourist reviews on TripAdvisor for the protection of the world heritage sites: Text mining

- approach. *Journal of Multidisciplinary Academic Tourism*, 6(1), 37–46. <https://doi.org/10.31822/jomat.876175>
- [33] Paolanti, M., Mancini, A., Frontoni, E., Felicetti, A., Marinelli, L., Marcheggiani, E., & Pierdicca, R. (2021). Tourism destination management using sentiment analysis and geo-location information: A deep learning approach. *Information Technology & Tourism*, 24(2021), 241–264. <https://doi.org/10.1007/s40558-021-00196-4>
- [34] Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), 143–157. <https://doi.org/10.1016/j.joi.2009.01.003>
- [35] Republic of Turkey Ministry of Culture and Tourism. (2021). *Kapadokya (Cappadocia)*. <https://tanitma.ktb.gov.tr/TR-22783/kapadokya.html>
- [36] Republic of Turkey Ministry of Culture and Tourism. (2020). *Turizm İstatistikleri (Tourism Statistics)*. <https://yigm.ktb.gov.tr/>
- [37] Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(23), 1–29. <https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0085-1>
- [38] Saura, J. R., Palos-Sanchez, P., & Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*, 11(3), 1–14. <https://doi.org/10.3390/su11030917>
- [39] Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2014, January 21-24). *Sentiment analysis; extracting decision-relevant knowledge from UGC* [Conference Session]. Information and Communication Technologies in Tourism, Dublin, Ireland. [https://www.researchgate.net/publication/260601169\\_Sentiment\\_Analysis\\_Extracting\\_Decision-Relevant\\_Knowledge\\_from\\_UGC](https://www.researchgate.net/publication/260601169_Sentiment_Analysis_Extracting_Decision-Relevant_Knowledge_from_UGC)
- [40] Schuckert, M., Liu, X., & Law, R. (2015). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. *International Journal of Hospitality Management*, 48(2015), 143–149. <https://doi.org/10.1016/j.ijhm.2014.12.007>
- [41] Sheldon, P. J. (2006). Tourism information technology. In L. Dwyer & P. Forsyth (Eds.), *International handbook on the economics of tourism* (pp.399–418). Edward Elgar Publishing. <https://www.elgaronline.com/display/9781843761044.00030.xml>
- [42] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- [43] Song, Q., Hu, W., & Xie, W. (2002). Robust support vector machine with bullet hole image classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 32(4), 440–448. <https://ieeexplore.ieee.org/document/1176893>
- [44] Statista. (2021a). *Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 1st Quarter 2019*. <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>
- [45] Statista. (2021b). *Number of User Reviews and Opinions on Tripadvisor Worldwide from 2014 to 2019*. <https://www.statista.com/statistics/684862/Tripadvisor-number-of-reviews/>
- [46] Ting, S. L., Ip, W. H., & Tsang, A. H. C. (2011). Is Naïve Bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications*, 5(3), 37–46. [https://www.researchgate.net/publication/266463703\\_Is\\_Naive\\_Bayes\\_a\\_Good\\_Classifier\\_for\\_Document\\_Classification](https://www.researchgate.net/publication/266463703_Is_Naive_Bayes_a_Good_Classifier_for_Document_Classification)
- [47] UNWTO. (2001). *eBusiness for tourism: Practical guidelines for destinations and businesses*. World Tourism Organization. <https://www.e-unwto.org/doi/epdf/10.18111/9789284404599?role=tab>
- [48] Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, 30(1), 123–127. <https://doi.org/10.1016/j.tourman.2008.04.008>
- [49] Visa, S., Ramsay, B., Ralescu, A., & VanDerKnapp, E. (2011, April 16-17). *Confusion Matrix-based Feature Selection* [Conference Session]. The 22nd Midwest Artificial Intelligence and Cognitive Science Conference, Ohio, United States of America. <https://openworks.wooster.edu/facpub/88>
- [50] Wei, L., Wei, B., & Wang, B. (2012). Text classification using support vector machine with mixture of kernel. *Journal of Software Engineering and Applications*, 5(12B), 55–58. <https://www.scirp.org/journal/paperinformation.aspx?paperid=26881>
- [51] Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management*, 31(2), 179–188. <https://doi.org/10.1016/j.tourman.2009.02.016>
- [52] Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58(2017), 51–65. <https://doi.org/10.1016/j.tourman.2016.10.001>
- [53] Xiang, Z., Gretzel, U., & Fesenmaier, D. R. (2008). Semantic representation of tourism on the internet. *Journal of Travel Research*, 47(4), 440–453. <https://doi.org/10.1177/0047287508326650>
- [54] Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert System with Applications*, 36(3), 6527–6535. <https://doi.org/10.1016/j.eswa.2008.07.035>
- [55] Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex & Intelligent Systems*, 6(2020), 621–634. <https://doi.org/10.1007/s40747-020-00155-2>
- [56] Yu, L., Porwal, A., Holden, E. –J., & Dentith, M. C. (2012). Towards automatic lithological classification from remote sensing data using support vector machines. *Computer & Geosciences*, 45(2012), 229–239. <https://doi.org/10.1016/j.cageo.2011.11.019>