Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University – Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/akufemubid

Araştırma Makalesi / Research Article
e-ISSN: 2149-3367
DOI: https://doi.org/10.35414/akufemubid.1537513
AKÜ FEMÜBİD **25** (2025) 035101 (497-509)
AKU J. Sci. Eng. **25** (2025) 035101 (497-509)

BERTurk-Based Sentiment Analysis on E-Commerce Multi Domain Product Reviews

Makale Bilgisi / Article Info
Alındı/Received: 22.08.2024
Kabul/Accepted: 05.12.2024
Yayımlandı/Published: 10.06.2025

Çok Alanlı E-Ticaret Ürün İncelemelerinde BERTurk Tabanlı Duygu Analizi

Bekir TEKE, Seda Nur YAZICI, Gulseren ZAMIR, Ali Bugrahan BUDAK, Isil KARABEY AKSAKALLI*

Erzurum Technical University, Faculty of Engineering and Architecture Dept. of Computer Engineering, Erzurum, Türkiye

Abstract

Product reviews on e-commerce platforms constitute an important source of information for customers' shopping processes. Learning about various product features and evaluating user experiences makes shopping more reliable and provides sellers with valuable customer satisfaction feedback. In order for sellers to make strategic decisions about their products, customer satisfaction and product feedback should be analyzed in detail. For this purpose, sentiment analysis methods were applied to the data to analyze the sentiment of the comments. In this study, sentiment analysis was performed using comments from the Trendyol e-commerce site. Our dataset consists of a total of 73392 data, retrieved from six different categories: Computer, Phone, Shoes, Clothing, Cosmetics, Sports and Outdoor through Selenium. The generated dataset is published in the Kaggle public database. Since the distribution of positive, negative and neutral labeled classes is unbalanced in the obtained data, a second dataset was created by applying a cluster-based undersampling method. After the preprocessing stage, these datasets were divided into 80% training data and 20% test data. As a result of the experiments, among the traditional machine learning models, Support Vector Machines (SVM) gave the highest accuracy rate with 89% (original) and 84% (undersampled) in both datasets, while the BERTurk model, one of the transformer-based models, was determined as the most successful model with an accuracy rate of 96% (original) and 93% (undersampled) compared to all methods.

Keywords: *Natural language processing; Sentiment analysis; Machine learning, BERTurk.*

Öz

E-ticaret platformlarındaki ürün yorumları, müşterilerin alışveriş süreçlerinde önemli bir bilgi kaynağı oluşturmaktadır. Ürünlerin çeşitli özellikleri hakkında bilgi edinmek ve kullanıcı deneyimlerini değerlendirmek, alışverişi daha güvenilir hale getirirken satıcılara da müşteri memnuniyeti konusunda değerli geri bildirimler sağlar. Satıcıların ürünleriyle ilgili stratejik kararlar alabilmesi için müşteri memnuniyeti ve ürünle ilgili geri bildirimlerin ayrıntılı bir şekilde analiz edilmesi gerekmektedir. Bu amaçla, yorumların duygu durumunu analiz etmek için veriler üzerinde duygu analizi yöntemleri uygulanmaktadır. Çalışmamızda, Trendyol e-ticaret sitesinin yorumları kullanılarak duygu analizi yapılmıştır. Veri setimiz, Selenium aracılığıyla Bilgisayar, Telefon, Ayakkabı, Giyim, Kozmetik, Spor ve Açık Hava olmak üzere altı farklı kategoriden veri çekilerek toplamda 73392 veriden oluşmaktadır. Oluşturulan veriseti Kaggle açık veritabanında yayınlanmıştır. Elde edilen verilerde pozitif, negatif ve nört etiketli sınıf dağılımları dengesiz olduğu için küme tabanlı örnek azaltma yöntemi uygulanarak ikinci bir veriseti oluşturulmuştur. Önişleme aşamasından sonra bu verisetlerinin %20'si test ve %80'i ise eğitim verisi olarak ayrılmıştır. Deneyler sonucunda geleneksel makine öğrenmesi modellerinden Destek Vektör Makineleri (DVM) her iki veri kümesinde de %89 (orijinal) ve %84 (undersampled) ile en yüksek doğruluk oranını verirken, transformatör tabanlı modellerden BERTurk modeli %96 (orijinal) ve %93 (undersampled) doğruluk oranı ile tüm yöntemlere göre en başarılı model olarak belirlenmiştir.

Anahtar Kelimeler: *Doğal dil işleme; Duygu analizi; Makine öğrenmesi; BERTurk.*

1. Introduction

Natural Language Processing (NLP) refers to the encoding of human senses, such as seeing, hearing, feeling, and life experiences through language (Amirhosseini and Kazemian, 2019). It offers a variety of methods that make emotions, thoughts, and behavioral patterns conscious and allow them to be developed in a goal-oriented and constructive manner. These methods enable modeling of the behavior of individuals or groups by analyzing successful performances and occasionally comparing unsuccessful actions. Modeling behavior allows predictions of the future in many fields, and these predictions contribute to the growing use of sentiment analysis (Anvar Shathik and Krishna Prasad, 2020). Sentiment analysis applications using NLP are especially critical for sellers and customers on e-commerce platforms. Studies have shown that sellers' sales strategies and customer satisfaction change significantly in a positive direction after sentiment analysis (Lyu and Choi 2020).

*Sorumlu Yazar/Corresponding Author: Isil KARABEY AKSAKALLI e-posta/e-mail: isil.karabey@erzurum.edu.tr

Sentiment analysis plays a critical role in NLP and aims to identify attitudes, evaluations, thoughts, opinions, or judgments about a particular topic. With the rise of social media platforms, the proliferation of user-generated unstructured text has made sentiment identification imperative for individuals, businesses, and governments. Sentiment analysis methods can be broadly categorized into two main groups: dictionary-based and machine learning-based approaches. While dictionary-based methods attempt to identify the emotional aspects of text documents by evaluating the semantic orientation of words and sentences, machine learning-based methods construct classification models using labeled datasets and employ supervised learning techniques (Kratzwald et al. 2018). Various deep learning architectures such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRUs), and long short-term memory (LSTM) networks, can be used to perform sentiment analysis. In addition, the representation of text documents plays a critical role in realizing NLP tasks based on machine learning. Although traditional bag-of-words representations ignore syntax and grammar rules, recent developments in word embedding representations have addressed the issues of high dimensionality and sparsity by representing texts through fixed-length vectors. These developments have led to significant advances in text representation. When the literature on sentiment analysis is reviewed, product reviews on e-commerce sites or social media posts are often used as data sets. In sentiment analysis studies, stages such as data collection, data preprocessing, sentiment classification, creation of a recommendation system, and evaluation are performed. In addition to traditional machine learning methods, BERT (Alaparthi and Mishra 2021) is one of the most widely used sentiment analysis methods. It is a word embedding model based on transformer architecture that bidirectionally reads word sequences in the input text. Being bidirectional means that a word learns its meaning from both the words before and after it. BERT tries to extract the correct meaning by reading both sides. It contains an attention mechanism that allocates words to their vectors based on nearby words (Ullah et al., 2023). It is a state-of-the-art (SOTA) language model that is highly bidirectional and pre-trained on a wide range of English Wikipedia texts (Prottasha et al. 2022). Thanks to BERT's bidirectional approach, the meaning of texts is understood more accurately and comprehensively. We can examine the way BERT works under two main headings named pre-training and fine-tuning. After masked language modeling and next sentence prediction parts in the pre-training phase, BERT is fine-tuned and customized for specific tasks. Smaller and specific datasets are used during the fine-tuning phase.

In this study, a dataset comprising six categories was created using reviews from a Turkish e-commerce platform named Trendyol. In the dataset, unnecessary punctuation marks, emojis, and meaningless words were removed in the preprocessing steps. Then, tokenization, stemming, and vectorization operations were applied to prepare data for the model. During the training process, the comments were labeled as positive, negative, or neutral, and the accuracy rate was ensured as a result of this labeling. As a result of the analysis, the BERTurk model obtained the highest accuracy rate. The primary contributions of this study are summarized as follows:

- Consumers' comments were categorized and publicly accessible.
- Emotional ambiguity of comments was reduced by applying cluster-base undersampling method.
- The system minimizes the loss of time that consumers make in choosing products based on product reviews.
- Sellers can more effectively analyze customer satisfaction and develop more planned and effective sales strategies based on this analysis.

The following sections of this study are organized as follows: In Section 2 describes previous studies and related articles. Section 3 provides preliminary information about the materials and methods used. Section 4 discusses the implementation stages of the proposed method in this study. Sentiment analysis is explained, and the methodology used is described. Section 5 discusses and evaluates the results obtained by the machine learning methods used and the pre-trained transformer models. Section 6 summarizes the overall study results.

2. Related Work

In recent years, researchers have been actively investigating the critical role of sentiment analysis in online product reviews. These studies have emphasized how sentiment analysis influences consumer decision-making processes and contributes to product quality improvement. As the volume of online data continues to grow exponentially, researchers have pointed out the increasing necessity for robust and efficient information extraction techniques. The application of sentiment analysis to recommendation systems has emerged as a promising avenue to enhance system performance by interpreting user feedback more accurately. Sentiment analysis, which seeks to determine the sentiments conveyed in reviews, can be executed at different granularity levels: document, sentence, or attribute. The

primary methodologies employed for sentiment analysis are categorized into three main approaches: lexicon-based techniques, machine learning-based techniques, and hybrid approaches combining both. Among these, machine learning methods have generally outperformed traditional lexicon-based techniques, particularly in text classification tasks. Sasikala and Lourdusamy (2020) developed two advanced techniques for sentiment analysis and performance prediction of online products: the Deep Learning Modified Neural Network (DLMNN) and the Adaptive Neuro-Fuzzy Inference System (IANFIS). These models operate within three analytical frameworks: rating-based (GB), content-based (CB), and collaborative (CLB) analysis. The DLMNN model demonstrated efficiency by reducing training time while enhancing classification accuracy. Conversely, the IANFIS system was designed to forecast product performance and anticipate consumer behavior trends. This approach incorporated essential preprocessing, feature extraction, feature selection, and review classification. Sentiment polarity was determined by calculating scores, enabling the classification of reviews as positive, negative, or neutral. For predictive analysis, weighted data metrics such as keyword frequency, word polarity, support, confidence, and entropy were utilized. Performance evaluations revealed that the CLB scenario achieved the highest accuracy, with a rate of 96.83% on a dataset of 5000 records.

In the study proposed by Zada and Albayrak (2023), Random Forest and AdaBoost ensemble learning techniques were used to compare user comments and user ratings on an online platform in Turkiye. First, 2237 comments were collected in a specific category. Basic natural language processing techniques were applied on the comments, the data was cleaned and analyzed. As a result of the experiments, the highest success was obtained using the Random Forest algorithm in the category with four stars (87.9%). As a result of the study, it was seen that user comments and ratings are not always compatible. For this reason, it is concluded that e-commerce platforms should integrate machine learning and natural language processing techniques into their systems while collecting user reviews. The integration of sentiment analysis with recommendation systems has the potential to greatly enhance recommendation quality by offering granular insights into user preferences and aversions. With a refined understanding of user needs, these systems can deliver more personalized and context-aware suggestions. Karabila et al. (2023) introduced a recommendation system leveraging the BERT model to analyze user sentiments expressed in comments. Their six-stage framework encompassed data collection, preprocessing, sentiment classification, recommendation generation, and evaluation. Comparative analysis of machine learning models used in the study revealed accuracy scores of 80.61% for Naïve Bayes, 81.45% for Decision Tree, 81.61% for Logistic Regression, and 80.90% for Support Vector Machine, whereas the BERT model achieved a notably higher accuracy of 91%.

Table 1. Comparison of state-of-the-art sentiment analysis in the literature.

Article	Dataset	Method	Performance
(Ullah et al., 2023)	Amazon product reviews	QBERT	F1-Macro: 0.91
(Ahmed and Wang, 2023)	SemEval 2014 & Amazon product reviews	Embedding CNN & BiLSTM using NGD	F1-Score: 0.81 Accuracy: 0.83
(Zhao et al., 2021)	Taobao,JD, Amazon and Ebay websites	LSIBA-ENN	Accuracy: 92.01%
(Prottasha et al., 2022)	Facebook, Twitter, and YouTube comments	Bangla-BERT+LSTM Word2vec	Accuracy: 94.15%
(Sasikala and Mary Immaculate Sheela, 2020)	Food review dataset	DLMNN, IANFIS	Accuracy: 93.77%
(Onan, 2021)	Twitter	CNN-LSTM	Accuracy: 93.85%
(Shankar et al., 2024)	Flipkart Products	Support Vector Machine	Accuracy: 85%
(Zada and Albayrak, 2023)	E-commerce product reviews	RandomForest	Accuracy: 87.9%
(Karabila et al, 2023)	Amazon Musical Instruments reviews	BERT	Accuracy: 91% F1-Score: 95%
This Study	**E-commerce product reviews**	**BERTurk**	**Accuracy: 96% F1-score: 96%**

Table 1 provides a comparative overview of state-of-the-art sentiment analysis methods from the literature, highlighting their datasets, methods, and performance metrics. For example, Ullah et al. (2023) employed QBERT on Amazon product reviews, achieving an F1-Macro score of 91%. Ahmed and Wang (2023) combined Embedding CNN and BiLSTM, obtaining an accuracy of 83% on datasets like SemEval 2014. Zhao et al. (2021) proposed a Local Search Improvised Bat Algorithm based Elman Neural Network (LSIBA-ENN) on data from platforms such as Taobao and Amazon, reporting an accuracy of 92.01%. Prottasha et al. (2022) demonstrated a 94.15% accuracy using a Bangla-BERT+LSTM model for sentiment analysis of Facebook, Twitter, and YouTube comments. Our study

stands out by employing a BERTurk model on e-commerce product reviews, yielding an accuracy and F1-score of 96%.

3. Materials and Methods

In this study, we compared the performance of machine learning models and pre-trained transformer-based models on 73392 data obtained via web scraping to perform sentiment analysis in product comments. The data preparation process included text conversion to lowercase letters, removal of punctuation and stop words, stemming, and TF-IDF vectorization. After these preprocessing steps, training and classification were applied. The machine learning models used include Support Vector Machine (SVM) (Suthaharan and Suthaharan, 2016a), Random Forest (RF) (Rigatti, 2017), Naive Bayes (NB) (Xu, 2018), Logistic Regression (LR) (Maalouf 2011), k-Nearest Neighbor (k-NN) (Peterson, 2009), Gradient Boosting (GB) (Natekin and Knoll, 2013), and decision tree (DT) (Suthaharan and Suthaharan, 2016b). Moreover, among the pre-trained transformer-based models, mBert (bert-base-multilingual-cased), BERTurk (bert-base-turkish-cased), XLNet, and DistilBERT were applied and compared. All materials and methods used in this study are explained in detail.

3.1 Data collection phase

In the data collection phase, a novel Turkish product reviews dataset on Trendyol e-commerce platform was created. To ensure both data quality and integrity, the data collection process is divided into several stages. These stages are described below:

Automatic Data Extraction: Using the Selenium library (Gundecha, 2015), automatic access to product review pages on the Trendyol platform was provided. By managing the Chrome driver via ChromeDriverManager (Sasikala and Mary Immaculate Sheela 2020; Wang et al. 2021), the relevant product pages were dynamically loaded and the page was scrolled by controlling the scroll bar so that all comments on the page were visible. This automation made the collection of large volumes of data efficient by providing access to all reviews of each product. This first step included a delay setting to account for the loading time between page scrolls. This ensured that no reviews were missed due to incomplete page loads.

HTML Parsing and Filtering: Once the visible page content was collected, the HTML content was parsed using the BeautifulSoup library and relevant information was extracted according to the HTML structure of the website (Shariff, 2019). Only the sections containing customer reviews were selected, excluding irrelevant text such as advertising content that was not included in the dataset.

Manual Data Labeling: After a total of 73392 comments were captured, each comment was manually checked and labeled into positive, negative or neutral categories for sentiment analysis. The labeling resulted in 60109 positive, 11867 negative and 1416 neutral samples. Manual labeling, performed by three anotators, was used to ensure consistency and reduce bias. Following specific guidelines, anotators labeled positive comments with expressions of satisfaction or high scores, negative comments with complaints or low scores, and neutral comments with mixed or objective statements. To maintain labeling quality, agreement checks were regularly performed between anotators to ensure consistency across emotion labels. The generated dataset can be accesed in Data Availabiliy section.

After the labeling process, the 15 most frequently used bigrams in positive, neutral and negative comments were extracted shown in Figures 1. The graphs in Figure 1 show that bigrams such as “tam”, “oldu”, “tam”, “beden” are commonly used in all class types.

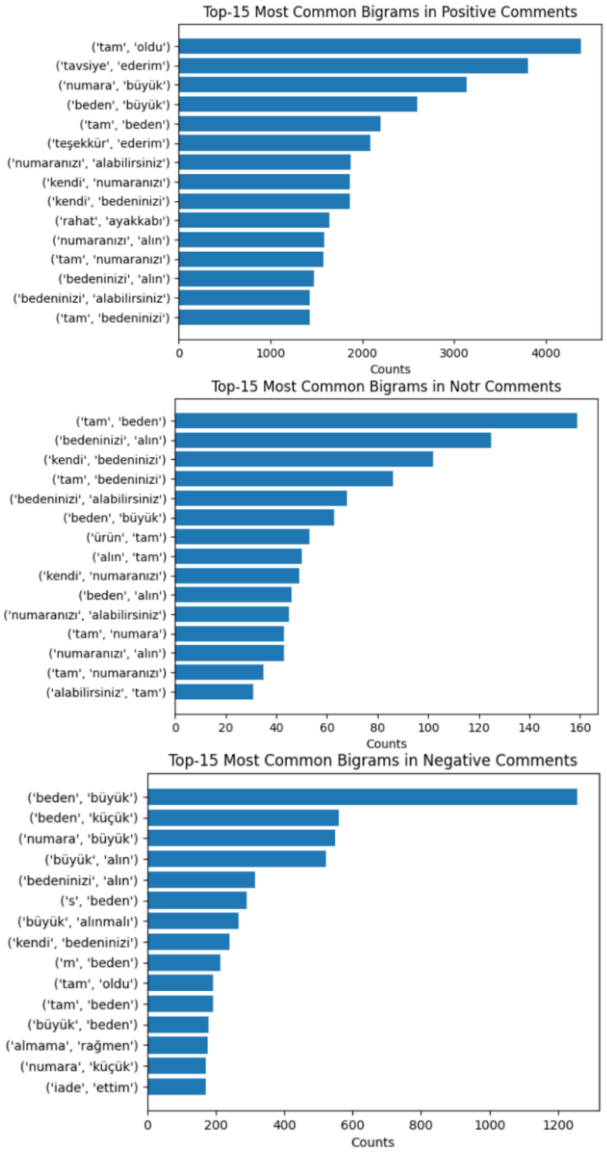

Figure 1. Top-15 Most Common Bigrams for Each Type of Comments

Data Quality Checks and Preprocessing: Before the dataset was finalized, data quality checks were performed to identify and remove duplicate, missing or irrelevant comments. The preprocessing step is one of the necessary steps to build models efficiently and accurately. The preprocessing step is designed to remove unimportant features from the text and select important features, revealing the distinctive features of the data. This process involves a series of operations that transform the raw text into a clean and structured format suitable for deep learning algorithms:

- *Lowercase conversion :* This step standardizes the text by converting all characters to lowercase. It ensures that words are treated the same regardless of case, which helps to identify patterns and semantics without case-related inconsistencies.
- *Removal of punctuation, numbers and emojis:* They do not carry semantic meaning within the sentence and can create noise in the data. Removing them helps clean up the text and makes it easier for the model to focus on and process meaningful content.
- *Removal of stopwords:* Stopwords are common words that do not add significant meaning to the text. Removing stopwords reduces the size of the text and improves the ability to identify patterns and related features in the data by shifting the focus to meaningful words. In the scope of the study, the Turkish stopwords of the nltk library as well as a manually defined stopword list were created and applied to the dataset.

The aim of this step is to prepare the data for artificial intelligence models by removing noise and emphasizing relevant information. After preprocessing, the dataset was found to be unbalanced in terms of class distribution. To avoid this, cluster-based undersampling method was applied to the original dataset. Cluster-based undersampling is an advanced undersampling method in contrast to the classical randomization-based undersampling method, which is applied when there are large imbalances in the data sets (Vigneron et al, 2016). This method first divides the samples from the majority class into various clusters. By selecting a few representative samples for each cluster, the total data set is reduced and made more balanced by ensuring that meaningful samples are selected from the majority class while preserving the diversity in the data set. This method aims to minimize information loss and provide a more meaningful representation of the majority class by selecting samples representing different regions in the dataset. In this dataset with a high number of positive classes, K-Means Clustering algorithm was used for clustering using feature vectors generated by TF-IDF (Lin et al. 2017). A certain number of samples were selected from each cluster and the dataset was balanced by reducing the total number of the positive class while preserving different examples of positive comments in the dataset. Thus, we obtained two different datasets on which we can measure performance, together with a new undersampling dataset with a total of 28283 samples, including 15000 positive, 11867 negative and 1416 neutral samples.

Figure 2 shows the general flowchart of data collection and web scraping processes. As a first step of the process, a cs file containing Trendyol reviews was read using the pandas library. Each line in this file contains a URL link to a product's reviews. Next, the selenium library was used to create a web driver that automatically controls the web browser; the tools required for this purpose include cabdriver and ChromeDriverManager (Sasikala and Mary Immaculate Sheela, 2020; Wang et al, 2021). A connection to the Trendyol website was established by launching the browser service. The browser was then opened to access the Trendyol website. To collect product reviews, we navigated to the relevant pages on Trendyol and scrolled the page so that all reviews were visible. After scrolling, the HTML content containing the reviews was extracted using the BeautifulSoup library, and the reviews were filtered according to a specific HTML structure (Shariff 2019). The resulting reviews were combined with other information, such as the product name, price, and review score, etc. and a single data frame was created. Finally, the data frame was exported to a CSV file containing reviews and other information about Trendyol products.

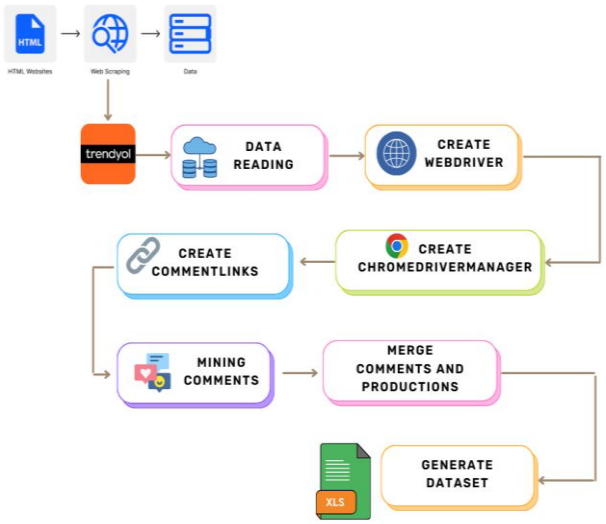

Figure 2. The general flowchart of web scraping and data collection phase

3.2 Sentiment Analysis

Today, one of the most important factors when shopping online is product reviews (Baubonienė and Gulevičiūtė 2015). There are many positive, negative, or neutral comments on the products we purchased. By analyzing the sentiments expressed in these comments, it is possible to evaluate the overall satisfaction of customers and make decisions accordingly. For sellers, developing product-related strategies, such as sales planning and material stock, depends on customer satisfaction (Jones et al. 2008). In this context, sellers must use sentiment analysis to understand customer satisfaction. Sentiment analysis in reviews is crucial for both customers and sellers to learn about a product and facilitate the shopping experience. Therefore, sentiment analysis on e-commerce platforms is a critical step to ensure an effective shopping experience. Sentiment analysis is a NLP technique that is used to automatically detect and classify emotions and feelings in texts. In the sentiment analysis process, the data cleaning phase after data collection is of great importance. This stage includes data preprocessing steps. When performing sentiment analysis on the commentary texts, elements that are not useful for analysis, such as numerical information, punctuation marks, and words without meaning, are removed. In addition, the entire text is converted to lower-case letters, and stemming is performed. After these preprocessing steps, the cleaned data are presented to the sentiment analysis algorithms. The results of the algorithm classified the emotions in the texts into specific categories. A general sentiment analysis flow diagram is presented in Figure 3.

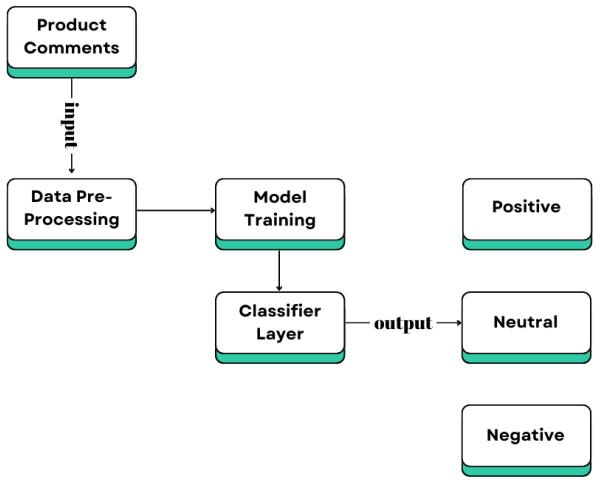

Figure 3. Fundamental sentiment analysis flowchart

There are various methods used in the field of sentiment analysis. These methods include algorithms such as Support Vector Machines (SVM), Naive Bayes (NB) and Long Short-Term Memory (LSTM), which are machine learning techniques. In addition, deep learning methods such as recurrent neural networks (RNN), convolutional neural networks (CNN) and Transformer-based models, especially BERT, are widely used for sentiment analysis. Deep learning models have the capacity to perform sentiment analysis with higher accuracy on large data sets. In this section, the sentiment analysis was examined in detail using both machine learning methods and deep learning approaches.

3.2.1. Sentiment analysis with Traditional Machine Learning Models

In this section, sentiment analysis was performed on Turkish product reviews using traditional machine learning methods. The goal of this analysis was to compare the sentiment classification performances of various machine learning algorithms. The dataset used included product reviews collected from the Trendyol e-commerce platform. The reviews were manually labeled as positive, negative, and neutral. Due to imbalanced data, we used undersampling method and created smaller dataset by dividing the samples in the majority class into various clusters. The two datasets were classified using machine learning methods. Table 2 and 3 shows the distribution of positive, negative, and neutral product reviews in the original and undersampled dataset resepectively.

Table 2. Number of instances according to the sentiments in the original dataset

Class	Number of instances
Positive	60109
Neutral	11867
Negative	1416

Table 3. Number of instances according to the sentiments in the undersampled dataset

Class	Number of instances
Positive	15000
Neutral	11867
Negative	1416

Figure 4 shows a flowchart of a basic sentiment analysis using machine learning models. After generating the dataset, preprocessing techniques were applied to a dataset of product reviews received from Trendyol customers. The aim was to highlight relevant information by removing noise from the data. These preprocessing steps are critical for ensuring that the deep learning model can learn effectively from the data and make accurate predictions. First, the text was standardized by converting all characters into lowercase letters. This process ensures that words are treated the same regardless of their uppercase/lowercase letters, thus eliminating inconsistencies between patterns and

meanings (Rathi et al. 2018). Removing punctuation marks helps clean up signs that create noise rather than meaningful content, which makes it easier for the model to focus on meaningful content. In addition, the text size was reduced by removing “stopwords” that did not meaningfully contribute, and focus was placed on meaningful words. This step improves the identification of patterns and important features in the data. The implementation of these preprocessing steps attempts to increase the performance and accuracy of deep learning models by improving the quality of the input data. The careful preparation of data is fundamental to the success of machine learning and deep learning tasks because it directly affects the model’s ability to learn and generalize from the data.

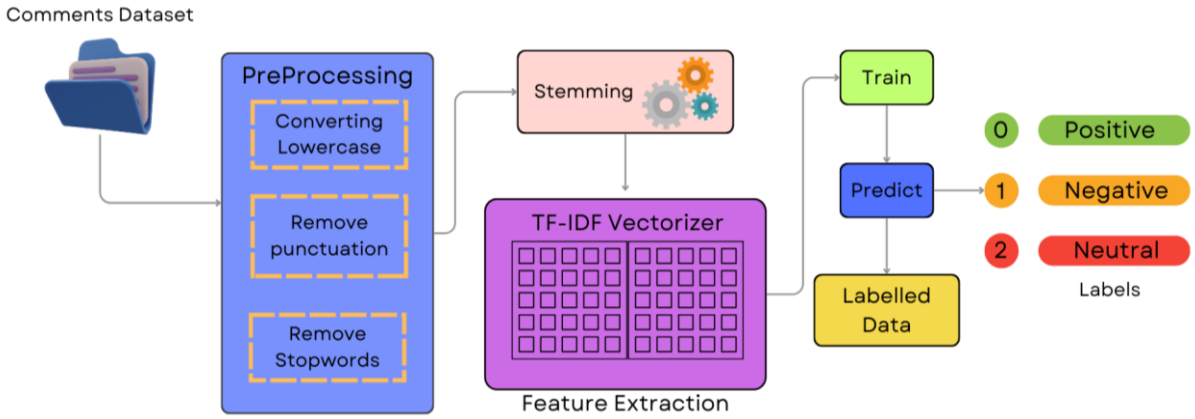

Figure 4. Flowchart of sentiment analysis with machine learning models

Stemming, a preprocessing method, is the process of reducing words to their root or basic form. This step is of great importance in the context of NLP because it reduces the dimensionality of data and allows words derived from the same root to be treated as a single feature. For example, the Turkish words "koşmak" (running) and "koşucu" (runner) can be reduced to the root "koş" (run). In our study, Snowball stemmer was used to effectively handle the morphological complexity of Turkish. This tool provides a convenient method for processing various root and suffix forms in the Turkish language. Then, we applied term frequency-inverse document frequency (TF-IDF) vectorization to the preprocessed and stemmed text data. TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is a statistical measure used to evaluate the importance of a word in a document against a collection of documents (Kubrusly and Valenotti 2024). This method determines the importance of a word by taking into account how often the word occurs in a given document (term frequency) and how common the word is in the entire collection (inverse document frequency). After data preprocessing and vectorization, the dataset was divided into training and testing sets. Various machine learning models used in this study were trained on training data and evaluated on test data. The models used included support vector machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), k-Nearest Neighbors (kNN), Gradient Boosting (GB), and Decision Trees (DT). Each model was trained on the same dataset to ensure fair comparison of performance.

The results of each model were compared based on performance metrics such as accuracy, precision, recall, and F1 score. The SVM model demonstrated superior performance with 89% and 84% accuracy rates, as shown in Table 4 and Table 7, demonstrating its effectiveness in emotion classification tasks. The evaluation metrics and complexity matrices provided detailed insight into the performance of each model, highlighting strengths and areas for improvement.

3.2.2. Sentiment analysis using BERT

In this study, in addition to traditional machine learning models, it was aimed to comprehensively evaluate the performance of sentiment analysis using the BERT model on Turkish product reviews. Within the scope of the study, it was aimed to compare the performances of both a transformer-based model such as BERT and classical machine learning algorithms such as SVM, Naive Bayes and Random Forest by training them on the same dataset. Reviews are divided into positive, neutral or negative categories and analysis processes are discussed in detail. In the research, data preparation, model training, evaluation and prediction processes are explained comprehensively. After the training was completed, the proposed models were tested on new probes and the results were comparatively discussed. Sentiment analysis is a critical area of NLP and focuses on accurately detecting and categorizing opinions in texts. This analysis is critical to understanding customer feedback, improving user experiences, and helping businesses make strategic

decisions. The BERT (Bidirectional Encoder Representations from Transformers) model used in the study stands out with its ability to analyze the meanings of words not only in context but also according to their relationships in the sentence. This feature makes BERT an ideal model for sentiment analysis, especially in morphologically rich languages such as Turkish. The model's ability to accurately capture the semantic relationships between words and cope with complex language structures increased its effectiveness in this project. As a result of this study, transformer-based models and traditional machine learning models were compared and the effectiveness of the modes was evaluated according to the performance metrics. The findings aim to guide businesses and researchers on which models can be used in sentiment analysis and how to get better results. Figure 5 shows the steps of sentiment analysis performed with the BERT model and the operation of the process in detail.

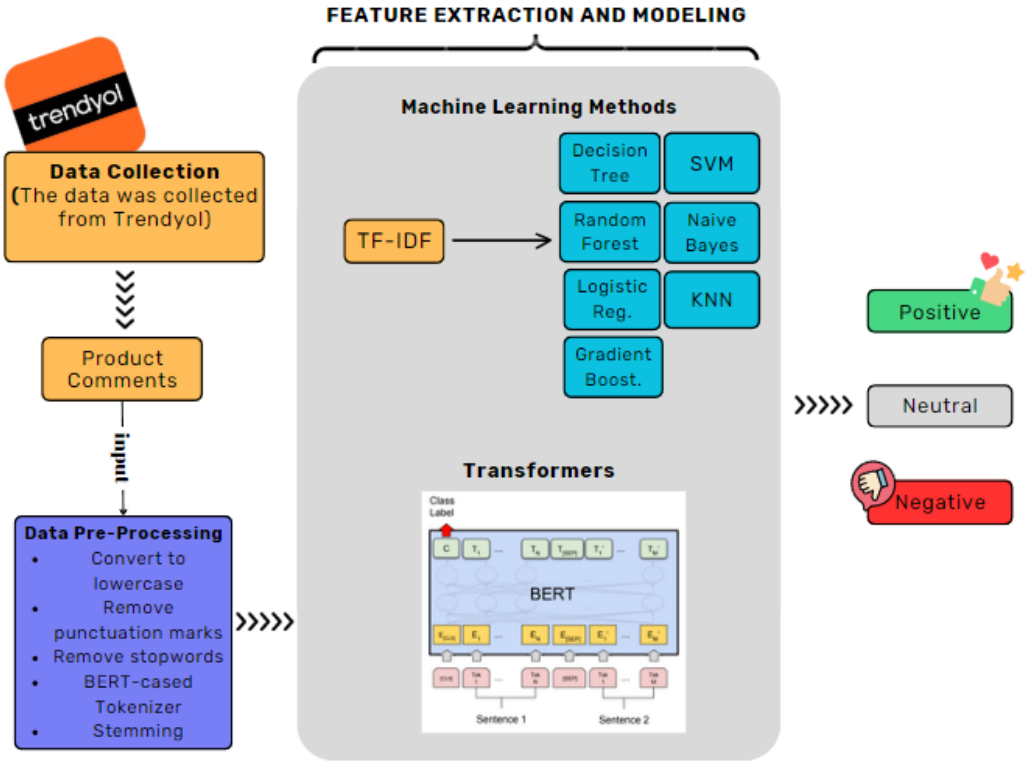

Figure 5. Flowchart of sentiment analysis with BERT model

First, the text data were processed using BERT tokenization to make them suitable for the input data of the BERT model. This process converts the texts into tokens that the BERT model can understand and organizes the data in the structure required by the model. During the tokenization process, special tokens ([CLS] and [SEP]) are added; the [CLS] token is used for classification tasks, and the [SEP] token indicates the distinction between sentences. While the tokens are converted to numerical values, attention masks are created that indicate which tokens the model should pay attention to. In these masks, the real tokens are denoted by 1, and the padding tokens are denoted by 0. For the model to process efficiently, all input sequences were filled to the same length, and short sequences were completed with zeros in this process. Processed tensors are prepared as inputs to the model, and the model receives these tensors. The model outputs were obtained in the form of raw scores (logits), which were converted to probabilities using the softmax function. These probabilities are used to determine which sentiment label has the highest accuracy. In the model training and classification phases, the BERT model was loaded and configured to classify the three emotional labels. If available, the model was transferred to a GPU, and the loss function was defined using the optimizer. The model was trained for several epochs, and during each epoch, the training loss was calculated and printed to the screen. The performance of the model was evaluated on the validation set, and the validation loss and accuracy were reported. After training, the performance of the model was analyzed in more detail using a classification report and confusion matrix; this analysis provides information about how well the model can distinguish between different emotion classes. The weights of the trained model were saved for future use. The trained model was loaded with a new dataset to

predict the sentiment of new comments, converted the comments into tokens, and then applied to generate emotional predictions. A function was defined to predict the sentiment of a given text, which converted the text into tokens, applied the model, and returned the probabilities along with the predicted sentiment. The comments were processed, and the predicted sentiments and probabilities of these sentiments were added to the dataset. The dataset was then saved in an Excel file. This study demonstrated the effectiveness of BERT in the sentiment analysis of Turkish product reviews. By carefully preparing the data, fine-tuning the model, and evaluating its performance, a robust sentiment analysis system was obtained. The data were classified into three categories: positive, negative, and neutral. The model's predictions can be used to obtain valuable information from customer feedback, which helps businesses in their decision-making processes and increases customer satisfaction.

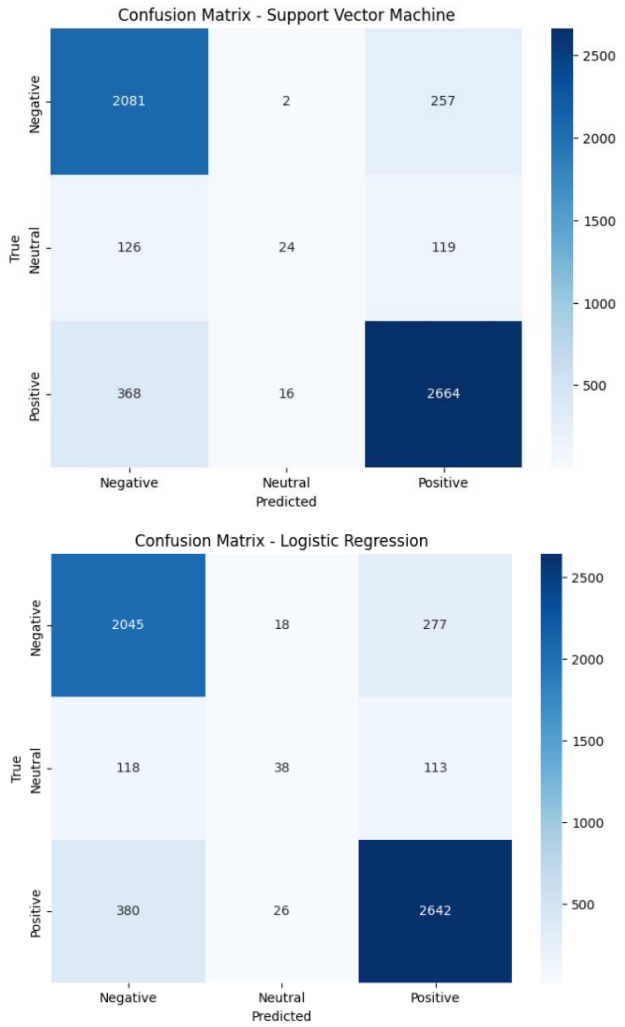

Figure 6. The confusion matrices obtained from the most successful traditional machine learning methods (SVM and LR) on cluster-base undersampled dataset

4. Experimental Results

In our dataset, 20% of the total comments were separated as test data, and the distinct remaining 80% were used as training data. This method allows a more reliable evaluation of model performance by dividing the dataset into training and testing stages. The training model was validated with 3-fold cross validation in machine learning methods. The confusion matrices obtained by the most successful traditional machine learning methods are presented in Figure 6.

The results obtained from the confusion matrix and calculated performance metrics demonstrate that the proposed model demonstrates high performance in some classes; however, improvement is required in other classes. As a result of evaluations conducted among traditional machine learning methods, the SVM model obtained the highest accuracy rate (84%) according to the confusion matrices. According to the confusion matrix obtained from the BERTurk model in Figure 7, the model performs better in distinguishing between neutral and positive emotions than traditional machine learning methods. 129 positive predictions were found as neutral, and 207 positive predictions are found as negative emotions. The other 11616 positive emotions were classified correctly. The number of false negative and false positive predictions was relatively low, indicating that the model generally provided high weighted average accuracy.

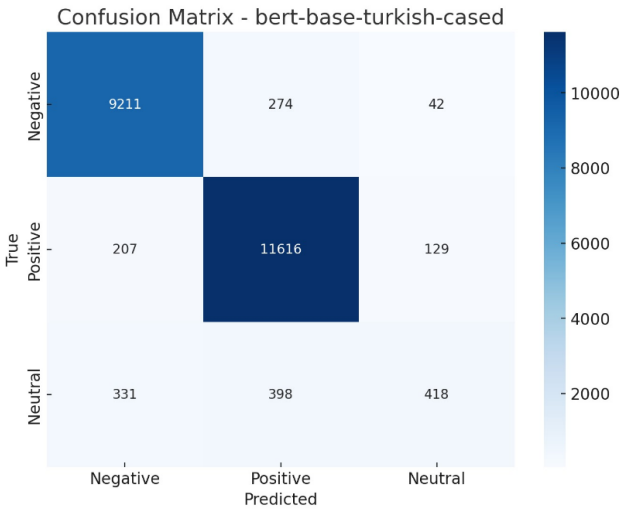

Figure 7. Confusion matrix obtained from BERTurk model on the cluster-base undersampled dataset

When we examine the sentiment analysis test results on original dataset in Table 4; although the accuracy rates of traditional machine learning models are promising, the differences in recall and F1-Score values demonstrate that these models are unsuccessful in predicting data belonging to certain classes. The reason for this failure

may be that these models are not sufficiently complex to analyze complicated data structures. The experimental results show that BerTurk, which is a pre-trained model with a multi-layer transformer architecture, produced higher results compare to the other transformer-based models and machine learning methods. Owing to the complex neural network structures, the model successfully processed long-distance dependencies and exhibited at least 7% higher accuracy than machine learning methods.

Table 4. Performance comparison of the machine learning models, mBERT, BERTurk, XLNet and DistilBERT on the original dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	**89**	**87**	**89**	**88**
RF	88	87	88	86
NB	83	83	83	77
LR	88	86	89	87
kNN	82	81	83	82
GB	85	84	86	83
DT	84	84	85	84
mBERT	95	95	95	95
BERTurk	**96**	**96**	**96**	**96**
XLNet	92	92	93	92
DistilBERT	83	83	83	83

In this study, promising results were obtained despite training on a very low number of steps by limiting the number of epochs to five to prevent overfitting. Table 5 presents the training and validation losses and validation accuracy depending on the number of steps. The gradual decrease in training loss and relative stability of validation loss and accuracy across epochs indicate that the model learned effectively but did not exhibit significant overfitting. Further increasing the number of epochs may lead to overfitting and reduce the ability of the model to generalize to unseen data. On the other hand, a smaller number of epochs may lead to underfitting, such as the model is not learning enough from the training data and performs poorly on both the training and validation sets. By training with five epochs, the model was allowed to learn from the training data for a sufficient period of time, and overtraining, which could lead to overfitting, was avoided. This approach allowed the model to maintain high validation accuracy throughout the training process.

Although the average precision and recall values of the BERTurk model are quite high compared to traditional machine learning models and other transformer-based models, the performance of the neutral comment analysis is lower than other classes due to the unbalanced distribution in our dataset. Class-based precision, recall and F1 score values are given in Table 6.

Table 5. Training and validation loss and validation accuracy per epoch using the BERTurk model on the original dataset

Epoch	Training Loss	Validation Loss	Validation Accuracy
1/5	0.28	0.30	0.91
2/5	0.20	0.25	0.92
3/5	0.15	0.24	0.92
4/5	0.12	0.20	0.92
5/5	0.10	0.19	0.92

Table 6. Precision, Recall, and F1-Score values for each class in the BERTurk model on the original test dataset

Class	Precision	Recall	F1 Score
Negative	0.92	0.93	0.93
Neutral	0.77	0.11	0.20
Positive	0.97	0.99	0.98

Table 7. Performance comparison of the machine learning models, mBERT, BERTurk, XLNet and DistilBERT on cluster-based undersampled dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	**84**	**83**	**84**	**83**
RF	82	81	83	82
NB	80	77	80	78
LR	83	82	84	83
kNN	72	73	73	73
GB	80	80	81	80
DT	77	77	78	77
mBERT	92	92	92	92
BERTurk	**93**	**93**	**94**	**93**
XLNet	88	88	88	88
DistilBERT	89	87	89	88

In order to increase the recall and F1 score values of the neutral labeled data in the original test dataset and to measure the performance of the algorithms on the balanced dataset, a more balanced dataset was created with the cluster-based undersampled method applied to the original dataset. Table 7 shows the performance of machine learning and transformer-based methods applied to the undersampled dataset. As seen in Table 7, the original dataset and cluster-based undersampled dataset showed similar performances in weighted average accuracy, precision, recall and F1 score values. SVM, one of the machine learning methods, outperformed the other machine learning models with an accuracy of 84%, while the BERTurk model outperformed all the methods performed in the experiment with an accuracy of 93% and an F1 score of 94%.

Table 8. Training and validation loss and validation accuracy per epoch using the BERTurk model on cluster-based undersampled dataset

Epoch	Training Loss	Validation Loss	Validation Accuracy
1/5	0.44	0.42	0.87
2/5	0.31	0.35	0.88
3/5	0.24	0.32	0.88
4/5	0.20	0.30	0.88
5/5	0.17	0.29	0.88

Table 9. Precision, Recall, and F1-Score values for each class in the BERTurk model on cluster-based undersampled test dataset

Class	Precision	Recall	F1 Score
Negative	0.94	0.97	0.96
Neutral	0.71	0.36	0.48
Positive	0.95	0.97	0.96

The training, validation losses and validation accuracy per epoch of the BERTurk model applied on the cluster-based undersampled dataset are given in Table 8, and the class-based performance of the BERTurk model on the test set is given in Table 9.

When the performance metrics of the neutral class obtained in Table 9 are compared with the metrics obtained from the original dataset shown in Table 6, an increase of 0.25 and 0.28 in recall and F1 score values is observed respectively. However, in terms of weighted average accuracy, precision, recall and F1 score values, the BERTurk model performed 3% better than the undersampled dataset due to the high number of positive values in the original dataset.

5. Discussion

In this study, we explored the efficacy of using the BERT model to analyze the sentiment of Turkish product reviews collected from the Trendyol e-commerce platform. The primary objective was to classify the sentiments expressed in reviews into positive, negative, or neutral categories, thereby providing valuable insights to help businesses improve customer satisfaction and product offerings.

Our experiments demonstrated that the BERTurk model outperformed traditional machine learning models and other transformer-based models in terms of accuracy, precision, recall, and F1-score (see Table 4 and Table 7). The traditional models, while effective to some extent, fell short in handling the complex structure of text data, which led to lower performance metrics. This underscores the limitations of classical approaches in capturing nuanced sentiments, particularly in a morphologically rich language like Turkish.

The BERTurk model's superior performance can be attributed to its ability to understand context and semantics through its bidirectional encoding mechanism. By fine-tuning the BERTurk model on our original dataset, we achieved significant improvements to the sentiment classification accuracy, reaching up to 96%. The confusion matrix (Figure 6) highlights the model's proficiency in distinguishing between positive and neutral sentiments, with high true positive and low false positive and false negative rates.

6. Conclusion

This study identifies positive comments that indicate customer satisfaction on e-commerce sites and helps understand the emotional states of customers by reading the comments. The dataset used to train the model was obtained through web scraping and included Turkish product reviews. The generated dataset was then semantically labeled with positive, negative, and neutral emotions. To make the data suitable for the model, the dataset was transformed and organized in the format required by the BERTurk model. The experimental results demonstrate that the proposed BERTurk model significantly exceeds traditional machine learning methods in terms of accuracy, precision, recall, and F1 score. The BERTurk model exhibited superior performance with a 96% and 93% accuracy rate in the original and undersampled dataset resepectively. When we consider the class-level performance, the results indicate that neutral labeled comments' classification is lower than the other classes. The reason why the precision, recall and F1 score values of the neutral labeled data in both datasets are lower than the other classes is that the words ("tam beden", "tam numara","kendi numaranızı", bedeninizi alabilirsiniz") in the comments with these labels are also included in the positive and negative comments. In conclusion, the experimental results have the potential to facilitate transactions between store owners and customers. Future research can improve the findings of this study. These may include expanding the dataset, integrating with more sophisticated recommendation systems, fine-tuning and optimizing the model, and developing real-time analysis systems. Overall, this study highlights the potential of the BERTurk model to revolutionize sentiment analysis in the e-commerce industry. The use of these techniques can provide businesses with deeper insights into customer feedback, thereby contributing to better decision making and increasing user satisfaction.

Declaration of Ethical Standards

The author declares that this study complies with Research and Publication Ethics.

Credit Authorship Contribution Statement

Author-1: Conceptualization, Methodology, Software, Validation

Author-2: Formal analysis, Investigation, Methodology, Resources, Validation, Data curation, Writing – original draft

Author-3: Formal analysis, Investigation, Resources, Data curation, Writing – original draft

Author-4: Formal analysis, Investigation, Resources, Data curation, Writing – original draft

Author-5: Data curation, Writing – original draft, Writing – review and editing, Visualization, Supervision, Validation, Project administration

Declaration of Competing Interest
There is no conflict of interest between the authors.

Data Availability
The generated datasets can be accessed via kaggle public repository:
Original dataset:
https://www.kaggle.com/datasets/aliburahanbudak/e-ticaret-yorumlar-duygu-etiketlenmi
Cluster-based undersampled dataset:
https://www.kaggle.com/datasets/aliburahanbudak/hazr-etiketli-e-ticaret-yorumlar-dengelenmi/data

Acknowledgement

This study is supported by the project fund number 1919B012319774 provided by the Scientific and Technological Research Council of Turkey (TÜBİTAK).

5. References

Ahmed, Z., and Wang, J., 2023. A fine-grained deep learning model using embedded-CNN with BiLSTM for exploiting product sentiments. *Alexandria Engineering Journal*, **65**, 731-747. https://doi.org/10.1016/j.aej.2022.10.037

Alaparthi, S., and Mishra, M., 2021. BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9, **2**, 118-126. https://doi.org/10.1057/s41270-021-00109-8

Amirhosseini, M. H., and Kazemian, H., 2019. Automating the process of identifying the preferred representational system in Neuro Linguistic Programming using Natural Language Processing. *Cognitive processing*, 20, **2**, 175-193.. https://doi.org/10.1007/s10339-019-00912-3

Anvar Shathik, J., and Krishna Prasad, K., 2020. A literature review on application of sentiment analysis using machine learning techniques. *Int J Appl Eng Manag Lett (IJAEML),* 4, **2**, 41-67. http://doi.org/10.5281/zenodo.3977576

Baubonienė, Ž., and Gulevičiūtė, G., 2015. E-commerce factors influencing consumers 'online shopping decision. https://doi.org/10.13165/ST-15-5-1-06

Gundecha, U., 2015. Selenium Testing Tools Cookbook. Packt Publishing Ltd. Birmingham, UK, 33-48

Jones, S. C., Knotts, T. L., and Udell, G. G., 2008. Market orientation for small manufacturing suppliers: The importance of product-related factors. *Journal of Business & Industrial Marketing*, 23, **7**, 443-453. https://doi.org/10.1108/08858620810901202

Karabila, I., Darraz, N., El-Ansari, A., Alami, N., & El Mallahi, M., 2023. Enhancing collaborative filtering-based recommender system using sentiment analysis. *Future Internet*, 15, **7**, 235. https://doi.org/10.3390/fi15070235

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., and Prendinger, H., 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision support systems*, **115**, 24-35. https://doi.org/10.1016/j.dss.2018.09.002

Kubrusly, J., and Valenotti, G. G. L., 2024. Comparison of document vectorization methods: a case study with textual data. *Sigmae*, 13, **1**, 79-90.

Lin, W. C., Tsai, C. F., Hu, Y. H., and Jhang, J. S., 2017. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, **409**, 17-26. https://doi.org/10.1016/j.ins.2017.05.008

Lyu, F., and Choi, J., 2020. The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews. *Sustainability*, 12, **11**, 4383. https://doi.org/10.3390/su12114383

Maalouf, M., 2011. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 3, **3**, 281-299.

Natekin, A., and Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7, **21**. https://doi.org/10.3389/fnbot.2013.00021

Onan, A., 2021. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and computation: Practice and experience*, 33, **23**, e5909. https://doi.org/10.1002/cpe.5909

Peterson, L. E. J. S. , 2009. K-nearest neighbor. 4, **2**, 1883. CURRICULUM VITAE. https://doi.org/10.4249/scholarpedia

Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., and Baz, M., 2022. Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors*, 22, **11**, 4157. https://doi.org/10.3390/s22114157

Rathi, M., Malik, A., Varshney, D., Sharma, R., and Mendiratta, S., 2018. *Sentiment analysis of tweets using machine learning approach*. In 2018 IEEE Eleventh international conference on contemporary computing (IC3), Noida, India, 1-3.

Rigatti, Steven J., 2017, Random forest. *Journal of Insurance Medicine*. **47**, 1, 31-39. https://doi.org/10.17849/insm-47-01-31-39.1

Sasikala, P., and Mary Immaculate Sheela, L., 2020. Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS. *Journal of Big Data*, 7, **1**, 33. https://doi.org/10.1186/s40537-020-00308-7

Shankar, A., Perumal, P., Subramanian, M., Ramu, N., Natesan, D., Kulkarni, V. R., and Stephan, T., 2024.

An intelligent recommendation system in e-commerce using ensemble learning. *Multimedia Tools and Applications*, 83, **16**, 48521-48537. https://doi.org/10.1007/s11042-023-17415-1

Shariff, S. M., 2019. Investigating Selenium Usage Challenges and Reducing the Performance Overhead of Selenium-Based Load Tests, Master Thesis, Kingston, Ontario, Canada, 100

Suthaharan, S., and Suthaharan, S., 2016. Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207-235. https://doi.org/10.1007/978-1-4899-7641-3_9

Suthaharan, S., & Suthaharan, S., 2016. Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269. https://doi.org/10.1007/978-1-4899-7641-3_10

Ullah, A., Khan, K., Khan, A., and Ullah, S., 2023. Understanding quality of products from customers' attitude using advanced machine learning methods. Computers, 12, **3**, 49. https://doi.org/10.3390/computers12030049

Vigneron, V., and Chen, H. 2016. A multi-scale seriation algorithm for clustering sparse imbalanced data: application to spike sorting. *Pattern Analysis and Applications*, **19**, 885-903. https://doi.org/10.1007/s10044-015-0458-2

Wang, X., Yi, G., and Wang, Y., 2021. *Automated Functional Testing of Search Engines using Metamorphic Testing*. In 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), Hainan, China, 22-29. https://doi.org/10.1109/QRS-C55045.2021.00014

Xu, S., 2018. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44, **1**, 48-59. https://doi.org/10.1177/0165551516677946

Zada, A. J. J., and Albayrak, A., 2023. Duygu Analizi ve Topluluk Öğrenmesi Yaklaşımları ile Kullanıcı Yorumlarının Analizi. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 11, **4**, 1725-1732. https://doi.org/10.29130/dubited.1102181

Zhao, H., Liu, Z., Yao, X., and Yang, Q., 2021. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing & Management*, 58, **5,** 102656. https://doi.org/10.1016/j.ipm.2021.102656