Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 05
Issue 02

December, 2024

Research Article

Review Mate: A Cutting-Edge Model for Analyzing the Sentiment of Online Customer Product Reviews using ML.NET

Özge CÖMERT[1], Nurcal YÜCEL[2]

[1] *Department of Transportation Services, Samsun University, 55000, Samsun, Turkey*
[2] *Department of Business Administration, Fırat University, 23000, Elazığ, Turkey*

ARTICLE INFO

Article history:
Received December 3, 2024
Revised December 12, 2024
Accepted December 17, 2024

Keywords:
Sentiment analysis,
Machine learning,
Classification,
ML.NET

ABSTRACT

E-commerce has become increasingly important in recent years due to several factors such as convenience, global reach, lower costs, personalization and uninterrupted access. In e-commerce, product reviews by customers can significantly impact purchasing behavior by providing social proof, establishing trust, aiding decision-making, improving search engine optimization, and increasing sales. Conducting an evaluation of the primary impacts of customer reviews on purchasing behavior through automated machine learning techniques has the potential to facilitate the advancement of diverse online business models. In this scope, we come with a new machine-learning model for evaluating customer sentiment based on product reviews. To this aim, a dataset consisting of 1000 positive and 1000 negative customer reviews was created by collecting publicly shared comments from online shopping websites serving in Turkey with a data collection tool developed by our research group. The model development was carried out on ML.NET, an open-source and cross-platform machine learning framework. In order to reach the most efficient model, a total of 36 machine learning models were explored for the solution of the problem within the scope of the experimental study. As a result, the model named Lbfgs Logistic Regression Binary was found to be the most efficient. The related model provided an accuracy rate of 94.76%. An API service called Review Mate has been developed to expand the potential impact of the proposed machine learning model and enable its use in different online business models. According to the findings, the proposed method outperforms the previous approach in terms of classification performance and also provides avenues for the discovery of new product ideas.

1. Introduction

Over the past few decades, the internet has revolutionized the way we live our lives and conduct business. E-commerce, in particular, has played a crucial role in this transformation. With the ability to shop from anywhere, at any time, e-commerce has expanded our reach to customers all around the world [1]. This increased access to a global marketplace has opened up countless opportunities for businesses to grow and succeed [2]. In addition, e-commerce has made it easier for consumers to find and purchase products that suit their needs and preferences. As technology continues to advance, the future of e-commerce looks bright, with even more potential for growth and innovation [3].

In order to stay competitive and meet the evolving needs of their customers, companies are increasingly focused on gathering and analyzing feedback. By

* Corresponding author
e-mail: ozge.comert@samsun.edu.tr
DOI: 10.55195/jscai.1595798

collecting customer comments and reviews, businesses gain valuable insights into the strengths and weaknesses of their products and services [4].

Armed with this information, companies can make strategic improvements, from minor tweaks to major overhauls, to better meet the needs of their target audience. Whether it's improving the functionality of a website, streamlining a checkout process, or introducing a new product line, companies that actively listen to their customers and incorporate their feedback into their operations are better positioned for success [5]. By prioritizing customer satisfaction, businesses can build brand loyalty, improve their reputation, and drive growth in the competitive marketplace.

Recently, it is seen that artificial intelligence-supported studies have been carried out to draw meaningful conclusions from customer comments [6]. These studies focus more on customer comments instead of using customer transaction databases containing demographic, psychographic, and purchasing behavior information traditionally used [7]. A framework called GSITK has been proposed for performing a wide range of sentiment analysis tasks such as dataset acquisition, text preprocessing, model design, and performance evaluation. It is intended for both researchers and practitioners and offers implementations of common tasks as well as facilitates the replication of previous sentiment models [8].

This study proposes an interpretable machine learning approach to customer segmentation for new product development, using online product reviews. Interpretable machine learning identifies non-linear relationships with high performance and transparency. Customer segmentation using the proposed approach outperforms a previous approach based on emotions. The findings demonstrate superior clustering performance and opportunities for new product concepts [9]. Research has been carried out to collect consumer feedback for airline service reviews and perform various forms of analysis on them. The study proposes a classification approach using machine learning techniques to identify positive and negative words or comments on a text-oriented database. Several ML algorithms such as Naive Bayes, Support Vector Machine (SVM), Decision Tree (DT), and Google BERT model are used to test the performance of sentiment analysis. In term of performance criteria such as accuracy, precision, recall, and F1-score, it is discovered that BERT outperforms other ML techniques [10].

A sentiment analysis was conducted using a substantial corpus of textual data from individuals commenting on a range of services and products. This analysis led to the formation of new datasets in Turkish, English, and Arabic, marking the first occasion on which comparative sentiment analysis has been conducted on texts in these three languages. Furthermore, the researchers were presented with a comprehensive study comparing the performances of both pre-trained language models and deep learning and machine learning models for Turkish, Arabic, and English. [11]. Other research has looked at customer satisfaction with baby products on Amazon. This was achieved through text mining and survey-based methodologies. The research model was developed based on factors derived from text mining, and a questionnaire was distributed and analyzed using Partial Least Squares Structural Equation Modelling (PLS-SEM). The results revealed that a number of dimensions influence customers' experiences of baby products. [12]. A different study proposes a CNN and LSTM-based supervised deep learning classifier for the multi-class analysis of sentiment in Bengali social media posts. These posts, which have been labelled according to their content, may be sexual, religious, political or acceptable. Six machine learning models have been developed as the basis for this classifier, with two different feature extraction techniques. These have been evaluated on a labelled dataset of 42,036 Facebook comments, with an accuracy of 85.80% achieved by the CLSTM architecture [13]. In a separate study, a dataset called Abusive Turkish Comments (ATC) has been proposed for the detection of abusive comments in Turkish on Instagram. In this study, sentence-level sensitivity annotation is performed. The performance of five well-known classifiers (Naïve Bayes, SVM, DT, RF and LR) and two re-weighted classifiers (i.e., Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost)) is evaluated in terms of F1-score, precision and recall. Ultimately, the findings indicated that the CNN model exhibited the most optimal performance on the oversampled ATC dataset [14]. Another study identifies sentiments in social media texts using machine learning. E-commerce website reviews and product reviews have been transformed into a tabular format for machine learning-based sentiment analysis. The reviews have been classified into three categories: positive, negative, and neutral. This classification was based on review scores. Turkish sentiment analysis models have been developed using SVM, RF, DT, LR, and k-nearest neighbors (KNN). The cross-validation results on independent test data from the same e-commerce website demonstrate that the SVM-based and RF-based sentiment analysis models outperform the others. [15]. A further study proposes the development of a new optimized machine learning (ML) algorithm, designated the Local Search Improved Bat Algorithm Based Elman Neural Network (LSIBA-ENN), for the analysis of sentiment in online product reviews. The Web Scrapping Tool (WST) is used to collect customer reviews from e-commerce websites. The web scraping data is preprocessed and then classified into positive, negative and neutral categories using LSIBA-ENN. This process demonstrated that LSIBA-ENN exhibited the most optimal performance [16]. In other related work, an innovative computational intelligence framework for the efficient prediction of customer review ratings is presented. Specifically, the proposed framework integrates Singular Value Decomposition (SVD), dimensionality reduction techniques, Fuzzy C-Means (FCM), and Adaptive Neuro-Fuzzy Inference System (ANFIS). The output of the proposed approach exhibited high accuracy. Additionally, the results indicated that in the case of large datasets, only a subset of the data is necessary for the construction of a system for predicting the evaluation scores of textual reviews [17].

In this study, we have tried to make our model robust by selecting the most effective methods and combining the best ideas from related works. We also believe that we have introduced innovations that have the potential to contribute to the field.

The main contributions of this study to the related field can be listed as follows:

(1) In this study, we created a large dataset using online customer product reviews on e-commerce sites.
(2) We made the dataset open-access and shared the dataset as a downloadable material in CSV format.
(3) We approached the light preprocessing steps differently from other studies on sentiment analysis in Turkish and took care to leave the comments as much as possible without affecting the raw data.
(4) Overall, we explored a total of 36 machine learning model within the scope of our study and selected the algorithm that produced the best results for distribution.
(5) We moved the proposed model to an API service and turned it into a software that can be used with zero configuration.
(6) We shared the service and AI model we developed in a GitHub repository as open access. We found that the model we proposed can be adapted to different business models and can be easily used to draw meaningful conclusions from product reviews.

The rest of the study is organized as follows: Section 2 summarizes the dataset and methodology. Section 3 presents the findings and provides a discussion. Finally, Section 4 presents the concluding remarks.

2. Material and Methods

2.1. Dataset

Within the scope of the study, collecting customer reviews is an essential step for model training. In order to accomplish this step, real customer reviews on the well-known websites hepsiburada.com, n11.com and trendyol.com in Turkey were taken into consideration.

A data collection tool called "*Product Pulse*" was developed to collect and process the customers' reviews. This tool was built on the ASP.NET Core MVC project template and uses the MSSQL database to provide persistent storage of data.

Table 1 Distribution of comments according to the source collected

	Positive	Negative	Total

hepsiburada.com	174	109	283
n11.com	253	705	958
trendyol.com	573	186	759
Total	1000	1000	2000

Table 1 shows the distribution of the data collected as a result of an extensive study. As can be seen, a total of 283 comments were received from hepsiburada.com, 958 comments from n11.com and 759 comments from trendyol.com respectively. A total of 2000 online customer product reviews, both positive and negative, were collected to be used in model training by paying attention to the balanced distribution of the data.

The data set produced within the scope of the research can be downloaded from the GitHub repository where experimental studies are shared.

2.2.Preprocessing

The proposed model is intended to work directly on customer reviews and contribute to the development of various online business models. For this reason, we make sure that customer reviews are as raw as possible. For this reason, in the preprocessing step, we remove spaces at the beginning and end of sentences. We eliminate phrases that use quotation marks. We eliminate start-of-line (\r) and end-of-line (\n) escape expressions and convert all characters to lower case before feature extraction.

Cleaning and standardizing text data before analysis is important for obtaining more accurate and reliable results. Removing spaces ensures that the text is more consistent and readable by eliminating unnecessary spaces at the beginning and end of sentences.

Removing quotation marks simplifies the text by recognizing that quotation marks in texts may be unnecessary for some analysis processes and could complicate the analysis process.

Removing newline escape expressions ensures that unwanted errors that may occur during text data processing are eliminated, resulting in the text being processed properly. Converting all characters to lowercase ensures that the analysis process is more consistent in cases where case sensitivity may be important in text analysis processes.

These preprocessing steps contribute to cleaning and standardizing text data before analysis, resulting in more accurate and reliable results [18].

2.3.Machine Learning

Typically, the process of building a model using machine learning algorithms involves a few basic steps such as problem statement, feature engineering, obtain dataset, feature extraction, model training and model evaluation [19].

Firstly, data needs to be collected and pre-processed to ensure that it is clean, complete and in the correct format [20]. Then, an appropriate machine learning algorithm must be selected based on the nature of the problem and the type of data being used. The algorithm must then be trained using a training dataset, which involves feeding it with inputs and expected outputs to enable the algorithm to learn patterns and relationships in the data [21].

After the algorithm has been trained, it can be tested using a separate test dataset to assess its accuracy and performance [22]. The model can be further refined and improved by changing parameters or using different algorithms, and it may be necessary to repeat the training and testing process several times to achieve the desired results.

After fine tuning and validating the model, it can be used to make predictions or generate insights based on new data inputs. The model may need to be continuously monitored and improved to ensure that it remains accurate and effective over time.

2.4.ML.NET

The Microsoft ML.NET framework was announced in May 2018. The framework is published on the GitHub platform as open access under the MIT license.

The Microsoft ML.NET library is already used in Chart Decision in Excel, Slide Design in Power Point, Windows Hello and Azure Machine Learning products.ML.NET is designed to be intuitive for experienced .NET developers. At the heart of ML.NET is the *MLContext* object, a singleton class like *AppContext*.

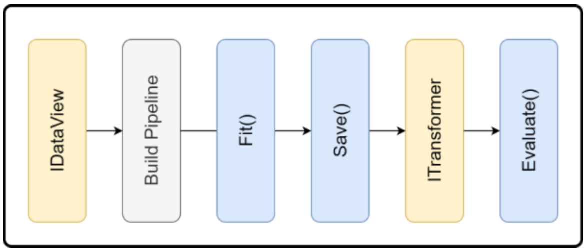


Figure 1 The high-level architecture of ML.NET

The *MLContext* object provides access to the entire catalog of trainers that ML.NET offers, such as anomaly detection, binary classification, clustering, regression and time series [23].

Fig. 1 shows the high-level architecture of ML.NET framework. Each block in the architecture can be summarized as follows: *IDataView* manages data loading, cleaning, and transfer in ML.NET, facilitating efficient and scalable operations. *BuildPipeline* creates sequential data processing and training chains, streamlining workflow construction. *Fit* trains datasets with specific algorithms to establish relationships between features and targets. Save stores trained models for easy retrieval and reuse. *ITransformer* enables result extraction and application, while *Evaluate* assesses model performance, guiding decision-making processes. Together, these components enhance ML.NET's usability, efficiency, and reliability in machine learning operations.

2.5. Proposed Model and Service

The general block diagram of the developed application is shown in Fig. 2. In the following sections, the details of the related blocks are explained in relation to the literature. Details of the first block related to data acquisition are already presented in Section 2.1.

As for *ReviewMateLib*, this is an ASP.NET Core library (*classlib*) project. Basically, the model data, training, evaluation, consumption and distribution of the generated model all take place within this library project. This project mainly consists of two main blocks. The first one performs the model training function, while the second one focuses on model consumption. In the model training phase, a pipeline for model creation, training and evaluation is built. The following steps are followed in the pipeline: In first step, *FeaturizeText* function is used to convert text data into numeric vectors. This function uses techniques such as bag-of-words model, n-gram features, TF-IDF criteria and similar techniques to make the text attributes suitable for model training. *This* function extracts a set of features for representing text data in vector space. This allows

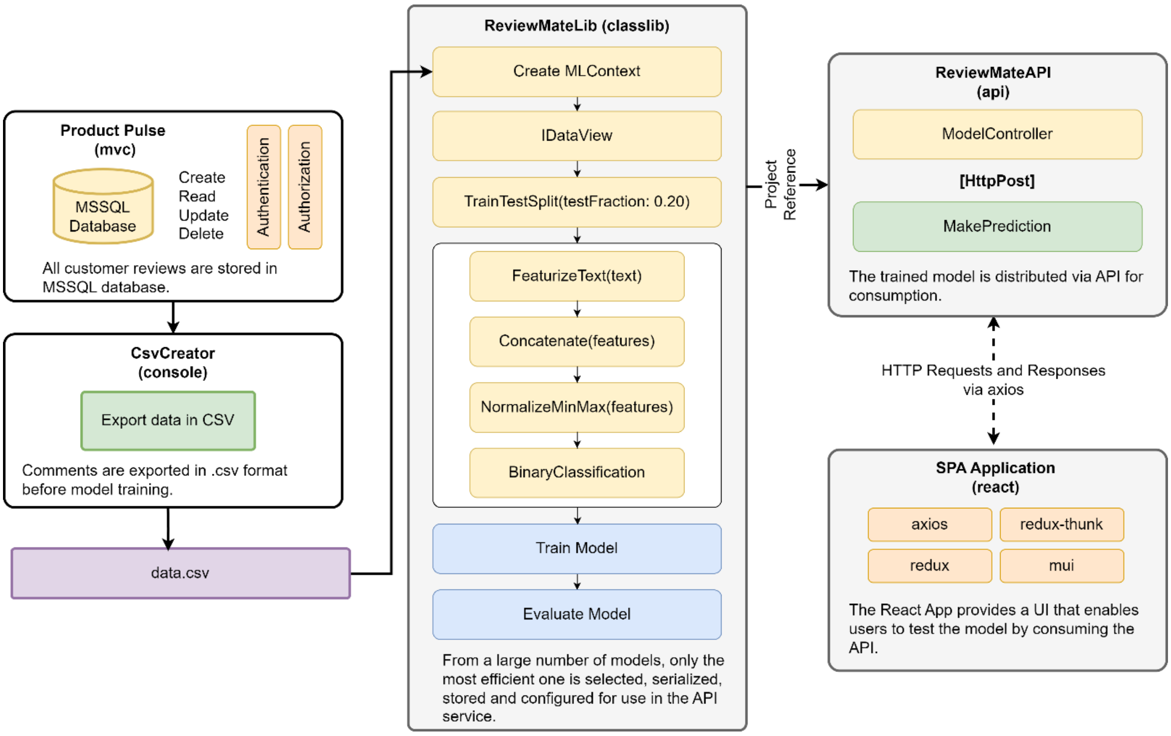


Figure 2 General block diagram of the study

text data to be represented by numeric values that can be used in machine learning algorithms. This function is a tool of ML.NET used in natural language processing (NLP) and is an important step to convert text data into numeric data. *Concatenate* in ML.NET performs a column concatenation operation and combines the specified columns into a single column.

Normalization is a process that brings the values in a dataset to a scale range. This process is used to balance the effects of variables with different scales, improve the performance of optimization algorithms, reduce the risk of overfitting, and increase the convergence speed of optimization. It ensures that the model learns more balanced and stable results. *NormalizeMinMax* takes a specific column of features and rescales all values to a specific range (usually [0,1] or [-1,1]). This function can help outputs perform better in a model. It is especially useful in cases where an un-normalized feature column has a larger range of values in terms of scale. This function can be used to pre-process data and can be combined with other transformations as part of an ML.NET pipeline. In Eq. (1), y represents the normalized value, which is between 0 and 1. x is the normalized feature.

$$y = \frac{(x - min)}{(max - min)} \tag{1}$$

In the classification step, binary classification models with different hyper-parameters are considered to obtain the most effective model.

Basically, four different machine learning algorithms are used for this purpose. These models are LbfgsLogisticRegression, FastTree, SdcaLogisticRegression, and LightGbm respectively.

Based on varying hyperparameter values, a total of 36 machine learning models are explored in this study. For a more concise and understandable presentation, only the Logistic Regression model that provides the best generalization performance is presented in the next section.

2.6. Logistic Regression

Logistic regression (LR) is a classification algorithm used particularly when the dependent variable (outcome) is categorical. It is commonly used in binary classification problems.

This method predicts the probability of the dependent variable belonging to a certain category using a linear combination of independent variables (predictors). Essentially, logistic regression classifies data appropriately using a sigmoid function called the logistic function. By fitting the model to the data, it predicts the probability of belonging to a specific class.

Linear regression is a type of linear model and it matches feature vector $\boldsymbol{x} \boldsymbol{\epsilon} \mathbb{R}^{\boldsymbol{n}}$to a scalar as described in Eq. (2).

$$p(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \tag{2}$$

Herein, $p(Y = 1 \mid X)$ represents the probability of Y being 1. e is the base of natural logarithm. $\beta_0, \beta_1, \dots \beta_n$ are the coefficients of the model and $X_1, X_2, \dots X_n$ are the independent variables.

The model training is completed in accordance with the mentioned steps. From this point on, the trained model can be utilized. In this context, the *CreatePredictEngine* function loads a trained model and creates a prediction engine to make predictions using this model.

The prediction engine takes input data and calculates outputs based on the model created during training.

The predictive engine uses only the memory required for the prediction function and therefore works efficiently on large datasets. In the final step, the *Predict* function executes the prediction function based on the inputs using *PredictEngine*.

2.7. Model Evaluating

Several metrics are used to evaluate the performance of classification models. These metrics include accuracy, precision, recall, F1 score, ROC curve and AUC.

These metrics are based on the classifications of true positive (TP), false positive (FP), true negative (TN) and false negative (FN). In our case, TP and TN represent the number of positive and negative reviews that were actually correctly detected, respectively.

Table 2 Performance metrics for evaluation classification models

Measure	Formula	Short Description
Accuracy (Acc)	$\frac{TP+TN}{TP+FP+FN+TN}$	In machine learning, accuracy is a performance metric used to measure the number of correctly classified instances out of the total instances in a dataset. It provides a quick and easy way to evaluate the effectiveness of a classification model.
Sensitivity (Se)	$\frac{TP}{TP+FN}$	Precision, also known as recall or true positive rate, measures the proportion of true positive cases correctly identified by a classification model. Precision is particularly important in applications where the cost of a false negative (missing a positive case) is high.
Specificity (Sp)	$\frac{TN}{TN+FP}$	The specificity is a performance measure used to assess the ability of a classification model to correctly identify negative examples. It is a complement to sensitivity, which measures the model's ability to correctly identify positive examples.
F1-Score	$\frac{2TN}{2TP+FP+FN}$	The F-score is a performance measure that combines precision and recall into a single value and provides a balanced measure of a model's accuracy. It is the harmonic mean of precision and recall, with 1 indicating excellent precision and recall and 0 indicating poor performance.

Similarly, FP and FN match the number of positive and negative reviews detected as false. The correct interpretation and use of these metrics ensures that the performance of the model is not misleading and provides guidance for model improvement [24]. Table 2 summarizes the several performance metrics derived from the confusion matrix.

2.8. Model Distributing

The ASP.NET Core API project template is used to build RESTful web services. This template enables developers to provide a presentation layer where they can create, read, update and delete data using the HTTP protocol [25]. It is typically used for sending and receiving data in JSON format. The template provides developers with various features in the API such as routing, input validation and CORS settings, and can be easily customized for additional features such as JWT authentication. ReviewMateAPI is used to allow the trained model to be easily consumed by different applications. The MakePrediction post function defined in the ModelController allows different applications to consume the related service [26].

2.9. Single Page Application

Single Page Application (SPA) applications are web applications where all pages are rendered in the browser and not by the server. This allows for a faster and more dynamic user experience. React is one of the popular JavaScript libraries used in the development of SPA applications [23]. React has a component-based structure and automatically manages changing data and updates the user interface. React also offers many great tools and features such as Virtual DOM, JSX, React Router, Context API, Redux. These features provide developers with a powerful toolset for building and managing SPAs. A React application including axios, redux, redux-thunk and Material UI (mui) libraries was developed from the scope of work to test our model live on a user interface.

Figure 3 Visualizing the essence of the dataset through a word cloud

3. Results and Discussion

As a result of the study, a dataset consisting of a total of 2000 online customer product reviews was created. In order to support similar studies and to ensure that the proposed algorithms can be easily compared, the dataset has been shared as open access in .csv format. In Fig. 3, the essence of the dataset is visualized through a word cloud. A word cloud graph is used as a visual representation of words in text data. The visibility of words indicates how frequently they are used in the text.

All experimental studies were carried out on a local workstation with Windows 11 64-bit operation system and Intel© Xeon© Gold 6132 CPU 2.60 GHz. The memory was 131 GB. Depending on the variation of the hyperparameter values, a total of 36 different models were explored to solve the problem. The most efficient models with values of hyperparameters are presented in Table 3.

Table 3 is important for presenting the results of the experimental study conducted. Designed to compare the performance of 36 different models using various hyperparameter values, this table provides a detailed description of the hardware and software environment used in the study. With this information, the table allows for the identification of the most effective models and clearly determines the hyperparameter values associated with these models. These details are important for interpreting the results of the study and guiding future research endeavors.

The hold-out validation technique was used instead of the k-fold cross-validation technique as there was sufficient data for model training. In this context, 80% of the data was reserved for the training set, while the remaining 20% was used as the test data set.

Table 3 Hyperparameters with values for the most efficient models

Models	Parameters	Values
LbfgsLogistic Regression	L1 Regularization	0.03125
	L2 Regularization	15.4891
	Tolerance	1E-8
	Enforce Non-Negativity	false
FastTree	Number of Leaves	20
	Number of Trees	100
	Minimum Example Count per Leaf	10
	Learning Rate	2
	Shrinkage	1
	Categorical Split	Categorical Split.Auto
	Feature Fraction	1
SdcaLogisticRegression	L1 Regularization	0.0
	L2 Regularization	1.0
	Maximum Number of Iterations	100
	Convergence Tolerance	0.0001
	Shuffle	True
	Optimization Tolerance	1E-8
LightGbm	Learning Rate	0.1
	Number of Leaves	20
	Number of Iterations	100
	Minimum Example Count per Leaf	20
	Maximum bin Count per Feature	256

In line with the aforementioned models and parameters, 80% of the dataset was used as the training set. The training stage was completed in 237.19 seconds and a total of 36 model discoveries were completed. Actually, these 36 models are basically based on four regression models given in The confusion matrices of the models regarding the classification results are given in Fig. 4.

The LbfgsLogisticRegression algorithm produced 94.76% of accuracy, 92.39% of sensitivity, 97.06% of specificity, and 94.55% of F1-score, respectively.

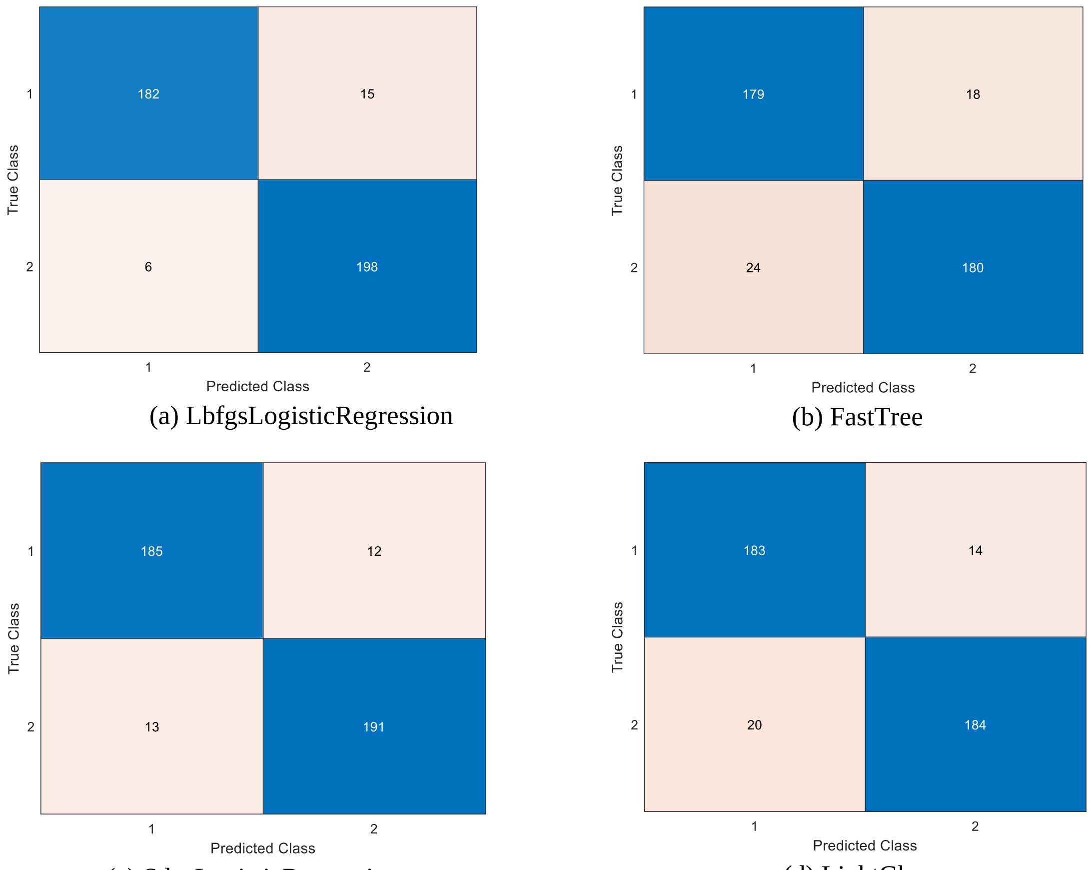
This model provided the best generalization performance. Similarly, the SdcaLogisticRegression algorithm provided satisfactory results. This algorithm yielded 93.77% of accuracy, 93.91% of sensitivity, 93.63% of specificity, and 93.67% F1-score. The results for other algorithms are given in Table 4.

business models. In this context, the results were considered promising and satisfactory.

When the related studies in the literature are examined, it is seen that almost all of the studies follow the steps of data collection, preprocessing, model development and model evaluation and the

Table 3 The classification results of the most efficient models with performance metrics

Trainer	**Acc**	**Se**	**Sp**	**F1-Score**	**AUC**
LbfgsLogisticRegression	0.9476	0.9239	0.9706	0.9455	0.9883
FastTree	0.8953	0.9086	0.8824	0.8950	0.9656
SdcaLogisticRegression	0.9377	0.9391	0.9363	0.9367	0.9833
LightGbm	0.9152	0.9289	0.9020	0.9150	0.9694



(a) LbfgsLogisticRegression

(b) FastTree

(c) SdcaLogisticRegression

(d) LightGbm

Figure 2 Confusion matrices of the models

The model development and evaluation studies revealed that the proposed model can automatically classify Turkish online customer product reviews with 94.76% of accuracy, 94.55% of F1-score and 0.9883 of AUC scores. The results show that the proposed model can be used in the innovative online studies are concluded in this manner. In this study, we developed an API service to extend the domain of our proposed model, to support online business development processes and to test the model with zero configuration. Moreover, we developed a SPA application with React JavaScript UI Framework to enable quick testing of the developed model and the model can be downloaded in the GitHub repository. The link of mentioned application is shared at end of the work.

In the other case, the Thunder Client VS Code Extension was used as an API testing tool. An HTTP Post Request was sent directly to the API service and the generated response was received. The service

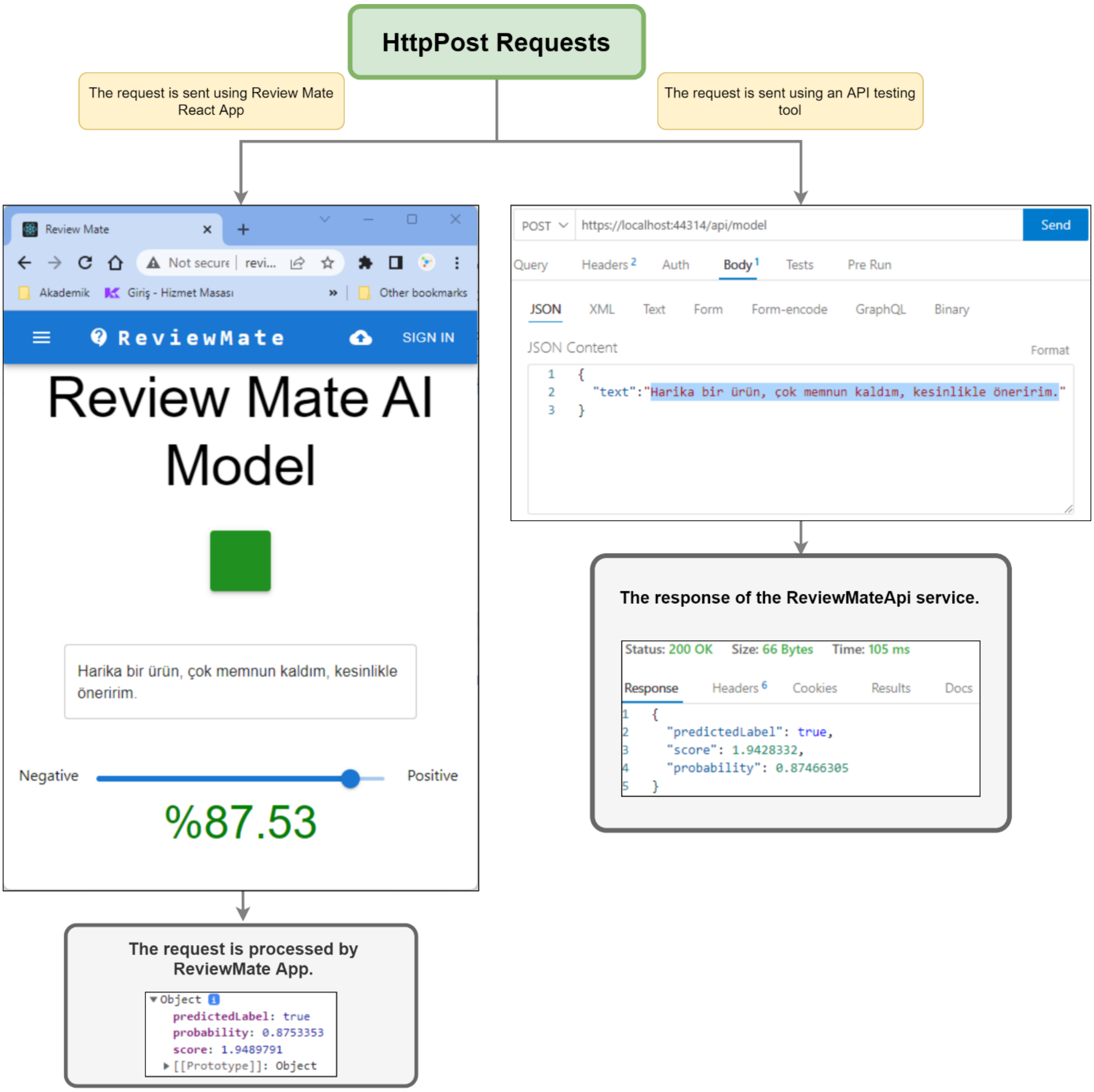


Figure 3 Analyzing sentiment of online customer product reviews with Review Mate Application

The overall diagram of the system designed in previous Fig. 2 is provided. The screenshot of the SPA application in this diagram is given in Fig. 5. Fig. 5 visualizes the processing of Http Post requests through the ReviewMate React App and ReviewMate API service, respectively. As seen in the figure, the ReviewMate React App performs sentiment analysis by reacting directly to customer reviews. At this point, the ReviewMate React App consumes API service and generates the responses simultaneously while the end user enters a comment.

response generated for requests includes three fields: predicted label, probability and score. The same comment was sent to the service as two separate requests and in both cases the requests were classified as positive. However, although the results are quite close, the responses are not equal. This can be considered normal in the context of how machine learning algorithms work.

In Table 5, the generated responses of the proposed model for customer product reviews randomly

Table 4 Randomly selected customer reviews and the responses of the ReviewMate API

Reviews	**Language**	**Translated to Turkish using Google Translate**	**Scale**	**Predicted Class**	**Probability**
Thanks for an excellent product.	English	mükemmel bir ürün için teşekkürler.	5*	+	0.9531
Great iPad, never had a problem with it but I wish I would have order it with the wifi and internet feature but it's a good product and it was amazing for gaming and school	English	Harika iPad, hiç sorun yaşamadım ama keşke onu wifi ve internet özelliğiyle sipariş etseydim ama iyi bir ürün ve oyun ve okul için harika	5*	+	0.7519
Entregou tudo certo, bem antes do prazo combinado.	Portuguese	Kararlaştırılan son tarihten çok önce her şeyi doğru bir şekilde teslim etti.	5*	+	0.8258
O produto veio com defeito, a bateria só carrega até 42%.	Portuguese	Ürün kusurlu geldi, pil sadece %42'ye kadar şarj oluyor.	2*	-	0.2227
The touchscreen went bad after a few days. Had to keep tapping and dragging before it worked. It continually asks if I want to upgrade, you have to log in all the time, even though I just did that. I couldn't just return it, I had to call in to explain that I did the obvious to try to fix it. Return experience horrible.	English	Dokunmatik ekran birkaç gün sonra bozuldu. Çalışmadan önce dokunup sürüklemeye devam etmem gerekiyordu. Sürekli olarak yükseltmek isteyip istemediğimi soruyor, ben bunu yapmış olmama rağmen her zaman oturum açmalısın. Öylece iade edemedim, düzeltmek için bariz olanı yaptığımı açıklamak için aramak zorunda kaldım. Geri dönüş deneyimi korkunç.	1*	-	0.1437
Great I pad with a lot of power! I love it	English	Çok fazla güçle harika bir yastıklama yapıyorum! Bayıldım	5*	+	0.8648
Calidad precio 10/10	Spanish	Kalite fiyatı 10/10	5*	+	0.5352
ste Ipad es una de las mejores compras que he podido haber hecho, escogí el modelo de 64 GB y es suficiente si lo que buscas es ver series, películas, navegar por la red o incluso jugar, corre perfectamente Genshin Impact, LOL Wild Rift, o Brawls Stars, la potencia es suficiente para un usuario que busca un medio de entretenimiento, a parte la batería dura todo el día, sumamente satisfecho con este Ipad.	Spanish	Bu Ipad, yapabileceğim en iyi satın alımlardan biri, 64 GB modelini seçtim ve aradığınız şey dizi, film izlemek, internette gezinmek ve hatta oyun oynamaksa yeterli, mükemmel çalışıyor Genshin Impact, LOL Wild Rift veya Brawls Stars, gücü eğlence aracı arayan bir kullanıcı için yeterli, ayrıca pili tüm gün gidiyor, bu iPad'den son derece memnun.	5*	+	0.7308
特殊な素材を使ってるのか？ゴム部分が臭く感じました。耐久性使いやすさは 100 点かな！	Japanese	Özel malzemeler mi kullanıyorsunuz? Kauçuk kısmın kokusunu alabiliyordum. Dayanıklılık ve kullanım kolaylığı 100 puan!	4*	+	0.1328
背面の立てかける部分が最初から開きにくく、数回開閉したら根本から折れてしまいました。プラスチックではなく金具か何かにした方が良いと思います	Japanese	Arkadaki parçayı baştan açmak zor oldu bir kaç kez açıp kapadıktan sonra tabandan kırıldı. Bence plastik yerine metal bağlantı parçaları veya başka bir şey kullanmak daha iyi	1*	-	0.1955

selected from Amazon.com source in different languages are examined. In this context, customer product reviews of variable length in English, Portuguese, Spanish and Japanese languages are considered.

While evaluating the reviews, the stars given by the customers were taken into account. The foreign language reviews were translated into Turkish with Google Translate and this translation was directly applied as input to the model. The class and probability value predicted by the proposed model are presented. As a result, the evaluation results of customer product reviews translated into Turkish from different languages were found to be satisfactory.

In Table 5, the generated responses of the proposed model for customer product reviews randomly selected from Amazon.com source in different languages are examined. In this context, customer product reviews of variable length in English, Portuguese, Spanish and Japanese languages are considered. While evaluating the reviews, the stars given by the customers were taken into account. The foreign language reviews were translated into Turkish with Google Translate and this translation was directly applied as input to the model. The class and probability value predicted by the proposed

Table 5 Comparison of related studies

Methods	**Description dataset**	**# of samples**	**# of classes**	**Availability**
Hybrid deep learning model, feature selection [27].	Prompt Cloud dataset	400.000 samples	5	Private dataset.
Multi attention and Bidirectional GSNP model (MA-BiGSNP) [30].	SST-2, IMDB, MR, Twitter, AirRecord	Over 6.000 samples	V	Public datasets
Machine learning techniques [28].	Hotes Amazon	Four different datasets with various numbers	V	Private datasets
Cluster based classification model [29].	iPod	1.811	2	Private dataset
A heterogeneous network model [31].	IMDB Yelp 2013 Yelp 2014	Over 78.000	V	Public datasets
Machine learning and deep learning models [11].	Movie, Game , Small appliances, Technological products, Large home appliances	100.000	3	Private dataset
Pretrained word embeddings and deep learnings [32].	Two different datasets produced	1.5 million samples	10	Private datasets
Machine Learning techniques [15].	hepsiburada	31.750 samples	3	Private dataset
Embedding based deep learning [18].	hepsiburada, n11, trendyol	2000 samples	2	Public dataset
This work, Lbfgs Logistic Regression model and Machine Learnig .NET.	hepsiburada, n11, trendyol	2000 samples	2	Public dataset

model are presented. As a result, the evaluation results of customer product reviews translated into Turkish from different languages were found to be satisfactory.

Table 6 provides a comprehensive comparison, taking into account similar studies in literature. It should be noted that conducting a direct one-to-one comparison may not be feasible due to variations in methods, samples, number of classes, and distribution differences in the datasets, as well as the utilization of different software and hardware resources. In this context, preferred environments for experimental studies, datasets used, sample sizes, availability of studies for open access, methods, and performance metrics have been considered.

The majority of datasets utilized for developing various sentiment analysis methodologies are proprietary and have restricted accessibility, posing a significant disadvantage for researchers [11], [27]–[29]. Additionally, the limited access to these datasets makes it challenging for different research groups to compare or replicate studies conducted on the same datasets. This circumstance may lead to potential knowledge loss in terms of developing novel methodologies and comparing existing techniques in the field of sentiment analysis. The availability of openly accessible and standardized datasets can contribute to the advancement of the field by granting researchers access to a broader spectrum of data. In this context, the dataset we have created has been shared openly for accessibility.

Another significant observation in related studies is the variable number of classes and sample sizes in the datasets. This variation makes it difficult to make comparisons and evaluate models effectively [15]–[17]. When the datasets contain different numbers of classes or different sample sizes, it becomes difficult to accurately evaluate the performance of sentiment analysis models. Furthermore, the inequality in class distribution across datasets challenges the process of comparing model performance across different studies. Researchers often face challenges in standardizing evaluation criteria and methodologies due to these differences, hindering the ability to draw meaningful conclusions from comparative analyses. As a result, achieving consensus or establishing best practices in sentiment analysis becomes more difficult in the presence of such dataset variability. We consider it imperative to increase the number of data samples in our experimental study compared to the monitored studies, but we accept that there is a limitation in this regard.

According to the findings obtained from the discussion table, it is evident that datasets obtained from various shopping portals (hotes.com, amazon.com, hepsiburada.com, etc.) were used. This observation indicates that the differences in the native languages of the users providing feedback, quality expectations and many other variable attributes may lead to expected differences in the generalization performance of the proposed models. In fact, reviews on different shopping portals are likely to vary in terms of cultural differences and shopping experiences. Therefore, it is crucial to consider this diversity when evaluating the generalization capabilities of sentiment analysis models based on data from different shopping platforms. These factors may affect the performance of the models and consequently, the variability in the usability of the proposed models across different platforms should be taken into account.

4. Conclusion

E-commerce is revolutionizing sales opportunities by enhancing the shopping experience of customers and expanding the horizons of businesses. Customer reviews have a crucial role in forming the purchasing decisions of potential customers, providing them with real insights into the quality of products and services offered by businesses. In this study, we created an openly accessible dataset of 2000 customer product reviews. For this purpose, we built a custom application called Product Pulse, which is designed to securely store customer reviews indefinitely. In the experimental phase, we performed a comprehensive exploration on the dataset with a total of 36 models and tested each model with varying configurations of hyperparameters based on four different core models. The analysis revealed highly satisfactory and promising results, providing a remarkable classification accuracy of 94.76%. To demonstrate the potential applicability of our proposed model in various online business frameworks and to facilitate its integration with minimal configuration, we created a Review Mate Web API service for model consumption. Furthermore, ReviewMate, a React-based application, was designed to serve as an interface between the API service and end users. We have made the datasets, services and applications from this extensive study available to other researchers through a private GitHub repository.

In our future work, we intend to keep leveraging the Product Pulse tool to collect customer reviews and openly disseminate datasets. In addition, our research efforts will focus on developing new models and tools to deal effectively with the challenges inherent in sentiment analysis based on customer reviews.

Data Availability

The dataset, services, and apps are publicly shared on the GitHub repository. Click here.

References

[1] M. Guha Majumder, S. Dutta Gupta, and J. Paul, “Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis,” *J. Bus. Res.*, vol. 150, pp. 147–164, 2022, doi: https://doi.org/10.1016/j.jbusres.2022.06.012.

[2] S.-H. Chin, C. Lu, P.-T. Ho, Y.-F. Shiao, and T.-J. Wu, “Commodity anti-counterfeiting decision in e-commerce trade based on machine learning and

Internet of Things," *Comput. Stand. Interfaces*, vol. 76, p. 103504, 2021, doi: https://doi.org/10.1016/j.csi.2020.103504.

[3] F. Liu, S. Liu, and G. Jiang, "Consumers' decision-making process in redeeming and sharing behaviors toward app-based mobile coupons in social commerce," *Int. J. Inf. Manage.*, vol. 67, p. 102550, 2022, doi: https://doi.org/10.1016/j.ijinfomgt.2022.102550.

[4] E. M. Mercha and H. Benbrahim, "Machine learning and deep learning for sentiment analysis across languages: A survey," *Neurocomputing*, vol. 531, pp. 195–216, 2023, doi: https://doi.org/10.1016/j.neucom.2023.02.015.

[5] N. Meilatinova, "Social commerce: Factors affecting customer repurchase and word-of-mouth intentions," *Int. J. Inf. Manage.*, vol. 57, p. 102300, 2021, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102300.

[6] X. Xu, R. Fan, D. Wang, Y. Wang, and Y. Wang, "The role of consumer reviews in e-commerce platform credit supervision: A signaling game model based on complex network," *Electron. Commer. Res. Appl.*, vol. 63, p. 101347, 2024, doi: https://doi.org/10.1016/j.elerap.2023.101347.

[7] D. Zhang, L. G. Pee, and L. Cui, "Artificial intelligence in E-commerce fulfillment: A case study of resource orchestration at Alibaba's Smart Warehouse," *Int. J. Inf. Manage.*, vol. 57, p. 102304, 2021, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102304.

[8] O. Araque, J. F. Sánchez-Rada, and C. A. Iglesias, "GSITK: A sentiment analysis framework for agile replication and development," *SoftwareX*, vol. 17, p. 100921, 2022, doi: https://doi.org/10.1016/j.softx.2021.100921.

[9] J. Joung and H. Kim, "Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews," *Int. J. Inf. Manage.*, vol. 70, p. 102641, 2023, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102641.

[10] A. Patel, P. Oza, and S. Agrawal, "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model," *Procedia Comput. Sci.*, vol. 218, pp. 2459–2467, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.221.

[11] P. Savci and B. Das, "Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 3, pp. 227–237, 2023, doi: https://doi.org/10.1016/j.jksuci.2023.02.017.

[12] M. Nilashi, R. A. Abumalloh, S. Samad, M. Alrizq, S. Alyami, and A. Alghamdi, "Analysis of customers' satisfaction with baby products: The moderating role of brand image," *J. Retail. Consum. Serv.*, vol. 73, p. 103334, 2023, doi: https://doi.org/10.1016/j.jretconser.2023.103334.

[13] R. Haque, N. Islam, M. Tasneem, and A. K. Das, "Multi-class sentiment classification on Bengali social media comments using machine learning," *Int. J. Cogn. Comput. Eng.*, vol. 4, pp. 21–35, 2023, doi: https://doi.org/10.1016/j.ijcce.2023.01.001.

[14] H. Karayiğit, Ç. İnan Acı, and A. Akdağlı, "Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods," *Expert Syst. Appl.*, vol. 174, p. 114802, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114802.

[15] M. Demircan, A. Seller, F. Abut, and M. F. Akay, "Developing Turkish sentiment analysis models using machine learning and e-commerce data," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 202–207, 2021, doi: https://doi.org/10.1016/j.ijcce.2021.11.003.

[16] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," *Inf. Process. Manag.*, vol. 58, no. 5, p. 102656, 2021, doi: https://doi.org/10.1016/j.ipm.2021.102656.

[17] G. Cosma and G. Acampora, "A computational intelligence approach to efficiently predicting review ratings in e-commerce," *Appl. Soft Comput.*, vol. 44, pp. 153–162, 2016, doi: https://doi.org/10.1016/j.asoc.2016.02.024.

[18] N. Yücel and Ö. Cömert, "Müşteri Duyarlılığını Keşfetmek İçin Yapay Zeka Destekli Analiz ile Çevrimiçi Ürün İncelemelerinden Anlamlı Bilgiler Elde Etme," *Fırat Üniversitesi Mühendislik Bilim. Derg.*, vol. 35, no. 2, pp. 679–690, 2023, doi: 10.35234/fumbd.1305932.

[19] M. Kilinc and C. Aydin, "Feature selection for Turkish Crowdfunding projects with using filtering and wrapping methods," *Electron. Commer. Res. Appl.*, vol. 62, p. 101340, 2023, doi: https://doi.org/10.1016/j.elerap.2023.101340.

[20] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[21] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1157–1182, 2003, doi: 10.1016/j.aca.2011.07.027.

[22] Z. Cömert and A. F. Kocamaz, "A Study of Artificial Neural Network Training Algorithms for

Classification of Cardiotocography Signals," *Bitlis Eren Univ. J. Sci. Technol.*, vol. 7, no. 2, pp. 93–103, 2017, doi: 10.17678/beuscitech.338085.

[23] Microsoft, "ML.NET An open source and cross-platform machine learning framework," 2024. https://dotnet.microsoft.com/en-us/apps/machinelearning-ai/ml-dotnet

[24] Z. Cömert, A. F. Kocamaz, and V. Subha, "Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment," *Comput. Biol. Med.*, vol. 99, pp. 85–97, Aug. 2018, doi: 10.1016/J.COMPBIOMED.2018.06.003.

[25] Z. Cömert and Ö. Cömert, "A Study of Technologies Used in Learning Management Systems and Evaluation of New Trend Algorithms," *Bitlis Eren Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, vol. 7. Bitlis Eren Üniversitesi, pp. 286–297, 2018.

[26] A. Sills, M. Ahmed, Dotthatcom.com, F. Boumphrey, and J. Ortiz, "Chapter 6 - Web Development Using XML and ASP.NET," A. Sills, M. Ahmed, Dotthatcom.com, F. Boumphrey, and J. B. T.-X. M. L. N. E. T. D. G. Ortiz, Eds., Rockland: Syngress, 2002, pp. 231–282. doi: https://doi.org/10.1016/B978-192899447-3/50009-3.

[27] P. Rasappan, M. Premkumar, G. Sinha, and K. Chandrasekaran, "Transforming sentiment analysis for e-commerce product reviews: Hybrid deep learning model with an innovative term weighting and feature selection," *Inf. Process. Manag.*, vol. 61, no. 3, p. 103654, 2024, doi: https://doi.org/10.1016/j.ipm.2024.103654.

[28] A. Yucel, A. Dag, A. Oztekin, and M. Carpenter, "A novel text analytic methodology for classification of product and service reviews," *J. Bus. Res.*, vol. 151, pp. 287–297, 2022, doi: https://doi.org/10.1016/j.jbusres.2022.06.062.

[29] P. Vijayaragavan, R. Ponnusamy, and M. Aramudhan, "An optimal support vector machine based classification model for sentimental analysis of online product reviews," *Futur. Gener. Comput. Syst.*, vol. 111, pp. 234–240, 2020, doi: https://doi.org/10.1016/j.future.2020.04.046.

[30] Y. Huang, X. Bai, Q. Liu, H. Peng, Q. Yang, and J. Wang, "Sentence-level sentiment classification based on multi-attention bidirectional gated spiking neural P systems," *Appl. Soft Comput.*, vol. 152, p. 111231, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111231.

[31] L. Gui, Y. Zhou, R. Xu, Y. He, and Q. Lu, "Learning representations from heterogeneous network for sentiment classification of product reviews," *Knowledge-Based Syst.*, vol. 124, pp. 34–45, 2017, doi: https://doi.org/10.1016/j.knosys.2017.02.030.

[32] M. Aydoğan and A. Karci, "Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification," *Phys. A Stat. Mech. its Appl.*, vol. 541, p. 123288, 2020, doi: https://doi.org/10.1016/j.physa.2019.123288.