applied sciences

MDPI

Article

Comparison of Machine Learning Models for Sentiment Analysis of Big Turkish Web-Based Data †

Cemile Gökçe Özmen [1,*] and Selim Gündüz [2,*]

[1] Department of Business, Faculty of Economics, Administrative and Social Sciences, Hasan Kalyoncu University, Gaziantep 27500, Turkey
[2] Department of Business Administration, Faculty of Business, Adana Alparslan Türkeş Science and Technology University, Adana 01250, Turkey
* Correspondence: cgokce.elkovan@hku.edu.tr (C.G.Ö.); sgunduz@atu.edu.tr (S.G.)
† This study is extracted from the PhD thesis titled "Business analytics with data mining: An investigation of web based data with sentiment analysis" conducted under the supervision of Selim Gündüz at Adana Alparslan Türkeş Science and Technology University, Institute of Graduate School.

Abstract: E-commerce sites have generated large amounts of unstructured data as they allow millions of users to generate product reviews. Thus, although there have been significant improvements in the characteristics of big data, such as speed and volume, developing various analysis techniques to monitor, understand, and extract useful information from this web-based data has become challenging. This study aims to analyze cosmetic products on a Turkish-based e-commerce website with sentiment analysis and to create a new domain-specific Turkish sentiment dictionary model with manual labeling. In the study, a Turkish sentiment dictionary consisting of 65,378 words was created by manually labeling 875,455 product reviews for 24 cosmetic brands sold on the Turkey-based trendyol e-commerce site, and sentiment analysis was performed using this dictionary. The dataset, divided into seven product groups, was analyzed using K-NN, SVM, DT, RF, and LR algorithms to address three classification problems. The algorithms were evaluated with comparative analysis using accuracy, precision, recall, and f-1 score metrics. SVM gave the highest performance result with over 93% accuracy, 92% precision, 93% recall, and a 91% f-1 score in all product groups. The dictionary model created for the cosmetics industry in the study helps businesses and researchers to use their resources more efficiently and save time by performing fast and low-cost analyses on large datasets of product reviews. Moreover, by analyzing customer feedback, brands can offer long-lasting and environmentally friendly products that align with customers' feelings. Thus, businesses have the opportunity to develop or improve products.

Keywords: machine learning; natural language processing; sentiment analysis

Academic Editor: Rui Araújo

Received: 27 January 2025
Revised: 14 February 2025
Accepted: 17 February 2025
Published: 21 February 2025

Citation: Özmen, C.G.; Gündüz, S. Comparison of Machine Learning Models for Sentiment Analysis of Big Turkish Web-Based Data. *Appl. Sci.* **2025**, *15*, 2297. https://doi.org/10.3390/app15052297

1. Introduction

While the development of big data has been ensured by the information age's provision of technology and internet opportunities to people, many areas have emerged where people, organizations, and businesses will benefit. With the rapid increase in the use of e-commerce sites and social networks, consumers have started to share their experiences, opinions, and feelings towards a product, service, or brand. This data, generated in the millions in the digital environment, has made it almost impossible for businesses and researchers to extract useful information with traditional methods. In the study, by utilizing the potential advantages of big data, a new dictionary specific to the field was created by identifying the

words, jargon, and forms of expression that consumers use while attributing and revealing their sentiments towards cosmetic products. This study, which analyzes web-based text data generated by users in the digital environment, is important in analyzing Turkish texts. In addition, the study reveals its originality because no other study uses Turkish product reviews as a dataset in the cosmetics industry.

There is a need for automated, fast, and easy methods to be used in business analytics, brand reputation, and strategy development from the large amount of data that consumers produce in order to make their voices heard by brands and businesses, make suggestions to other customers, or share their experiences. Sentiment analysis and various machine learning methods are among the research tools used to meet these needs. This method is used to extract people's sentiments from web-based data. It was born as an extension of data mining using natural language processing techniques, and it is important in natural language processing, data and text mining, and big data. Sentiment analysis, mainly performed by processing, analyzing, and interpreting textual data, is an important tool in detecting people's emotions and predicting them for the future. This approach is important and offers advantages for individuals, businesses, various institutions, and even governments. Within the scope of the objectives of the thesis study, the following research questions were addressed:

RQ1: Can sentiment analysis be performed with a large Turkish textual dataset?

RQ2: Can a sentiment dictionary with high accuracy and performance be created based on Turkish product reviews?

RQ3: How do different machine learning algorithms evaluate sentiment analysis performance on Turkish cosmetic product reviews?

Businesses that benefit from sentiment analysis have advantages such as understanding customer behavior, following market trends, or learning about their competitors' situation. It is a suitable technology for perceiving and analyzing a consumer's behavior [1]. In addition to saving business time, it provides an understanding of what customers expect from the business and its products [2]. With sentiment analysis, subjective information in the source material is recognized, extracted, and characterized through contextual manipulation of the text. This helps many businesses to understand the social sensitivity of their products or services.

Thus, sentiment analysis enables early detection of negative emotional expressions, managing crises, and enabling customer relationship management to take practical actions in negative customer experiences. It determines an entity's attitude towards the text's overall contextual polarity and detects emotional state or emotional communication [3,4]. One of the most widely used strategies in sentiment analysis is machine learning. Machine learning is an approach that largely overlaps with statistics, and many evaluations are empirical [5]. With machine learning, a model is built with data using a good and helpful approach and is used to solve real-world problems. It has evolved from efforts to discover whether computers can learn to mimic the human brain to a discipline that generates fundamental computational theories in multidisciplinary fields such as statistics, information theory, philosophy, artificial intelligence, psychology, and neurobiology [6].

In this study, sentiment analysis was used to evaluate customer reviews for 24 different cosmetic brands on Trendyol, one of the largest e-commerce sites based in Turkey. Manual tagging created a domain-specific normalized dictionary model consisting of 65,378 words, and the model's performance was evaluated with various machine learning approaches. In the study flow, the sentiment analysis studies carried out in the cosmetics sector in the international literature were summarized. The materials and methods used were summa-

rized step by step. The classification performances of the machine learning algorithms were compared, then the findings were presented, and a discussion was carried out.

2. Literature Review

This study focuses on meaningful information extraction and model building with sentiment analysis and machine learning approaches from big data consisting of Turkish product reviews in the cosmetics industry. Various models have been constructed and applied in studies conducted in the cosmetics industry using the sentiment analysis research method.

Tran et al. performed aspect-based sentiment analysis on a dataset containing 16,227 Vietnamese reviews of lipstick products. This study evaluated model performance by comparing single-task and multi-task learning with a deep learning approach. They achieved a success rate of 98.09% with the BİGRU + Conv1D model. In addition, they determined that the classification performance of the proposed model specific to the Vietnamese language works with a 91.01% f-1 score success rate [7].

Salsabila and Sibaroni conducted a dimension-based sentiment analysis study using a support vector machine algorithm. Semantic similarity and TF-IDF weighting were added to cosmetic product reviews in Indonesia. In the study, which includes a dataset of sunscreen, tonic, serum, essence, scrubs, and exfoliating product groups, price, packing, and scent dimensions were extracted, and the model's performance was evaluated for each aspect. From the test results, the accuracy rates were 93%, 92%, and 86% for the price, packing, and scent aspects, respectively [8].

Guen and Juyoung conducted a sentiment analysis study on consumer reviews of BB cream products and compared competing products. The researchers compared the competitors with datasets including 4335 reviews of 42 products in the Korean product group and 6001 reviews of 194 products in the other product group. They used the discriminative features of words using word frequency analysis, LSA, and L-LDA. In the last stage, the proposed model was supported by a t-test, and the study concluded that Korean products are more advantageous in competition because they have a higher average [9].

Clara et al. performed aspect-based sentiment analysis with 5053 Indonesian product reviews in toners, serums and essences, scrubs and exfoliators, and sun protection cosmetics. This study used TF-IDF and n-gram feature extraction methods to extract price, packing, and fragrance dimensions, and classification performances were evaluated with the random forest algorithm. This study obtained 90.48% accuracy, 87.27% precision, 70.13% recall, and 71.77% f-1 score values [10].

Fadly et al. conducted a sentiment analysis study using a dataset of 9899 English-language product reviews of skin care products from Drunk Elephant, Origins, Belief, Laneige, and Glam Glow brands. Performing a comparative analysis with naïve Bayes, KNN, SVM, decision tree, and deep learning algorithms, the researchers achieved the best performance with deep learning and decision tree, with an accuracy close to 80% and an f-1 score of 60% [11].

Park performed sentiment analysis using a dataset of 35,617 customer reviews of 26 top global cosmetic brands and evaluated relative customer satisfaction through TF-IDF analysis. It is predicted that the approach proposed in the study can be applied to realize or improve customer satisfaction with cosmetic brands [12].

Jaehun et al. developed a model that classifies 110,097 product reviews of the top 26 cosmetic brands by sentiment analysis and reveals sentiment-attributed features by applying latent Dirichlet allocation [13].

Hung and Chao presented a new approach for sentiment classification from 12,000 Chinese cosmetic product reviews (WOMs). The authors created a dictionary model consisting of feature words and contextual words, evaluated the model's performance using SVM, J48 decision tree, and multilayer perceptron, and achieved a high success rate [14].

Romadhony et al. conducted a sentiment analysis study with a dataset of more than 700,000 Indonesian cosmetic product reviews, addressing the triple classification problem. They applied multinomial naïve Bayes, support vector machine, LSTM, and BILSTM to evaluate sentiment classification performance and proposed a domain-specific model for the Indonesian language. When the classification performances were evaluated with the study, the best performance result was obtained with BILSTM, and the prediction with MNB worked more successfully than SVM [15].

3. Materials and Methods

Although language models cannot be used in Turkish sentiment analysis as in English, all methods in the literature are suitable for use [16]. In this study, where the Turkish sentiment analysis model was created, the process flow of the proposal is presented in Figure 1. This proposal evaluated customer feedback from a dataset consisting of a large amount of cosmetic product reviews and measured the classification performance with machine learning approaches.

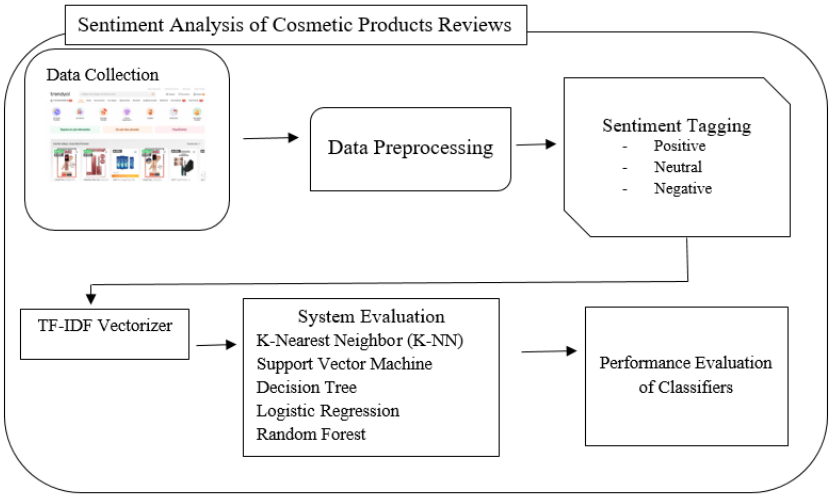

Figure 1. Turkish model process flow.

After the data collection and creation of the dataset, the dataset was made suitable for analysis with data preprocessing. The positive, negative, and neutral rates of the reviews of each cosmetic brand were calculated with manual emotion labeling. After the digitization process with the TF-IDF vectorizer, the dataset was separated into 80% training and 20% test data, and the system evaluation was performed by calculating the classification performance using machine learning approaches.

3.1. Data Collection

In order to apply the analyses of the study, the first step, data collection, was carried out. The data were collected by downloading from trendyol.com, an e-commerce website based in Turkey. Trendyol.com is an e-commerce site that serves clothing, accessories, electronics, home and life, supermarkets, cosmetics, and shoes and bags and is frequently preferred by users in Turkey. Trendyol.com, which also has a mobile application, allows customers to shop whenever and wherever they want and to generate product reviews.

In order to create the dataset, the top 50 global cosmetics brands published in 2023 by the Brand Finance consultancy firm in the UK were targeted. Among these brands, product groups sold on trendyol.com with at least 400 product reviews were selected. In addition, since the study focuses on women's cosmetic product groups, male care product brands in the list were not included in the study. Thus, 981,456 product reviews from product groups belonging to 24 different brands were obtained by data scraping and saved in the csv format. In this study, where the text mining method will be used, only the comments containing emojis or emoticons were deleted and cleaned. Finally, a dataset consisting of 875,445 product comments was prepared. Details on the dataset are presented in Table 1.

Table 1. Dataset.

Product Group	Brand	Reviews	Total Reviews
Skin Make-Up Products	Estée Lauder	2.985	54.579
	Garnier	1.123	
	M.A.C.	7.045	
	L'Oreal	27.225	
	Maybelline	16.201	
Eye Make-Up Products	Estée Lauder	823	78.345
	Lancôme	4.893	
	L'Oreal	38.577	
	M.A.C.	4.131	
	Maybelline	29.921	
Lip Make-Up Products	Clinique	7.750	62.729
	L'Oreal	2.080	
	M.A.C.	27.279	
	Maybelline	25.619	
Skin Care Products	Clinique	17.016	313.080
	Estée Lauder	10.036	
	Garnier	48.734	
	La Roche-Posay	97.661	
	Lancôme	2.723	
	L'Oreal	37.268	
	Kiehl's	11.742	
	Neutrogena	3.241	
	Nivea	47.491	
	Vichy	2.890	
	Yves Rocher	34.278	
Hair Care Products	Clear	9.272	158.505
	Dove	6.293	
	Elseve	66.461	
	Head&Shoulders	6.715	
	Herbal Essences	2.103	
	L'Oreal	5.094	
	Pantene	19.190	
	Vichy	18.800	
	Yves Rocher	24.577	
Body Care Products	AXE	1.359	181.256
	Dove	15.108	
	Garnier	13.825	
	Neutrogena	11.927	
	Nivea	55.162	
	Old Spice	829	
	Palmolive	15.364	
Perfume	Lancôme	6.624	26.699
	Oriflame	16.733	
	Yves Rocher	3.342	

3.2. Data Preprocessing

Data preprocessing is correcting internet-based data, noisy data such as incorrect words, slang and swear words, and abbreviations, keeping only the usable parts of the data, and discarding and filtering unwanted or unimportant parts [17]. It is an important step to be performed before applying NLP techniques in any language [18]. Data preprocessing raw data with polarity but that are not yet processed are susceptible to redundancy and inconsistency, and these raw data are preprocessed to improve the data quality [19].

3.3. Sentiment Tagging

The sentiment tagging stage was performed to analyze the sentiment of the preprocessed data. Based on the fact that the study consists of Turkish texts, sentiment expressions vary according to the context, and product reviews in the cosmetics industry are created in a specific context. The sentiment tagging process was performed manually and created a sentiment dictionary. In addition, although sentiment analysis has been carried out in various contexts in Turkish texts, no study has been found in the context of Turkish product reviews in the cosmetics industry.

Sentiment labeling was conducted according to word level and phrase level. Positive words/phrases were given a score of +1, negative words/phrases were given a score of −1, and words/phrases that did not contain any emotion were given a score of 0. In addition, the net score was determined according to the sum of the positive or negative sentiments in a sentence.

3.4. TF-IDF Vectorizer

Term frequency inverse document frequency score is based on the observation that important terms in a document have higher frequencies in that document and lower frequencies in the whole dataset [20]. It is a weighting mechanism to emphasize words that tend to be one of two classes [21]. It is used as a weighting element during information retrieval and provides a ratio with the appearance of a particular term in the document [22]. Term frequency is defined as TF, which estimates the number of times a term occurs in the text, and inverse document frequency IDF, which estimates how important a word is [23].

Term frequency measures how often a word occurs in a document and is calculated by dividing the number of times a word appears in a document by the total number of words in that document. Inverse document frequency is calculated as the logarithm of the number of documents in the corpus divided by the number of documents in which a particular word occurs. It measures how important a word is [24]. The IDF becomes zero if a term exists in all documents [25].

3.5. System Evaluation

At this study stage, the success of manually labeled data performed with sentiment analysis was tested with various machine learning algorithms. The success of the proposed model in the positive, neutral, and negative classifications was evaluated with K-NN, SVM, DT, RF, and LR.

Sentiment analysis is a real-time computational technique [26] that tries to understand and explain sentiments by analyzing large amounts of text data in a way that helps people make decisions [27]. As part of AI technology, sentiment analysis is a meaningful analysis method that obtains text sentiment trends and polarity [28]. It is located at the intersection of natural language processing and large-scale data mining [29]. It identifies, classifies, and evaluates human sentiments' polarity in subjective documents or texts.

One of the most widely used strategies in sentiment analysis is machine learning. Machine learning is an approach that largely overlaps with statistics, and most evaluations

are empirical [5]. With machine learning, a model is created with data using a good and helpful approach and is used to solve real-world problems [6]. Machine learning algorithms learn the characteristics of categories from a set of classified text and use them to sort documents into predefined categories after building an automatic classifier [30]. It is a part of artificial intelligence that provides the benefit of automatically extracting information from texts and concepts without direct and explicit programming [31]. Thus, the machine learning approach is a branch of artificial intelligence that develops techniques that allow computers to learn and aims to create generalizable programs from unstructured data [32].

K-NN is a non-parametric method used to perform robust classification or regression, especially for sentiment analysis [33], and is applied to estimate the probabilities of unseen co-occurrences [34]. The classification method selects K-nearest neighbors in training documents and classifies an unannotated document according to these K-neighbors [35]. In the training phase, only the feature vectors and categories of the training set are stored [30]. It performs classification by assigning unlabeled observations to the class of the most similar labeled samples [36]. In the algorithm, the 'K' value indicates the number of neighbors used in the model's estimation; however, in this study, K = 5 was used by default. Thus, class estimation was performed by looking at the 5 closest neighbors of the model.

SVM is used to perform subjective and objective classification and polarity classification applications [37]. The algorithm aims to select a hyperplane that maximizes the margin between the closest examples of two different classes [38] or can perform a classification that can discriminate between a certain amount of data from the training set [39]. It tries to find the best separation and classification between the data using support vectors [40]. The most appropriate hyperplane to be selected with the algorithm separates the data into two groups [41]. It finds a hyperplane separating two labeled classes with a maximum margin, which is shown as the distance between the two hyperplanes [42]. In this study, kernel = rbf (radial basis function) was used by default and a hyperparameter that determines the tolerance to error was used with the value C = 1.0.

Decision tree learning is one of the predictive modeling methods related to machine learning and is frequently used in data mining [43]. The essence of a decision tree is to learn from supervised data with association logic [44]. It generally represents and visualizes a decomposition of the combination of constraints of instances on feature values [45]. As a decision-making technology, the decision tree uses a tree diagram to represent the expected value of each decision [46]. The decision tree is a tree structure for manually categorizing training documents by generating true/false queries. In its structure, leaves represent the document category, while branches represent the combination of features leading to these categories [30]. Each non-leaf node is labeled with an attribute, which asks a question about the input sample and creates a branch for possible answers. Possible answers are determined by one of the answers from branches [40]. Thus, classification is obtained using the path from the root to the leaf node [47]. In the DT codes applied in the study, the depth of the tree is defined as unlimited (max_depth = none), thus allowing the model to learn the dataset.

RF is an algorithm based on ensemble learning. Through ensemble learning, where different types of algorithms or the same algorithm are combined multiple times, RF combines multiple algorithms of the same type. It combines multiple decision trees and thus creates a forest of trees [48]. This algorithm provides high classification accuracy and can determine variable importance. In this study, 100 trees were used by default (N_estimators = 100), and each tree was trained with the dataset to increase the generalization ability of the model.

LR is a discriminative classifier that allows the estimation of independent variables [49]. Independent variables are one or more than one; logistic regression is based on algorithms

that determine the output and result [50]. LR consists of extracting a set of weighted features from the input, multiplying each feature by the weight and then summing [51]. In the LR model used in the study, the iteration process was performed by determining the value of max_iter = 1000, thus determining the stopping criterion and allowing the model to learn better.

The use of these algorithms was preferred due to the high performance accuracies of supervised machine learning approaches and the creation of a trained model with manual tagging, which constitutes one of the foundations of the study. The suitability of the system for each of the product categories and groups that constitute the study's dataset was tested with a confusion matrix and evaluation metrics. The evaluation metrics used were accuracy, precision, recall, and f-1 score.

The confusion matrix is a matrix in which the classification model results are compared with the actual results, as shown in Table 2 [52]. If the model results classified in a text are the same as the actual result, it is considered correctly classified [53].

Table 2. Confusion matrix for sentiment classification.

Predicted	**Actual**	
	Positive Sentiment	**Negative Sentiment**
Positive Sentiment	True Positive (TP)	False Positive (FP)
Negative Sentiment	False Negative (FN)	True Negative (TN)

Accuracy calculates the percentage of the total predicted quantity that is correctly predicted [54]. With evaluation metrics, performance is calculated based on the accuracy in the test set [55].

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \quad (1)$$

TN/TP is the number of reviews where both humans and computers agree that they belong to class C; FP/FP is the number of reviews where humans classify the reviews as belonging to class C, but the classifier classifies the inaccuracies as not belonging to class C [56]. Precision, recall, and f-1 score calculations are performed with the following equations [57]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F-1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

While precision measures the accuracy of a classifier, recall is a tool that measures the model's prediction accuracy for the class, i.e., the ability to find all positive units in the dataset [58]. F-measure is the weighted harmonic mean of precision and recall [59].

4. Results

This section presents the experimental results of sentiment analysis on a sizeable textual dataset of cosmetic product reviews in Turkish. The analyses were performed using the Python 3.11. Jupyter Notebook software language.

In the first stage of the study, the dataset detailed in Table 1 was created, and data preprocessing was performed. It is an important step in performing the analyses of the study and obtaining accurate performance results. The large dataset used in the study caused various spelling errors, emoticons, emojis, unrelated words, or conjunctions to be frequently included in the texts due to the structure of natural language and

subjective textual data. For this reason, the data preprocessing stages performed in the study are as follows:

Step 1. Cleaning: In the data preprocessing stage, first, special characters, emojis, emoticons, punctuation marks, and extra spaces that cause noise and inconsistency in the raw data were removed. Thus, expressions that would not make sense in the product reviews and could negatively affect the analysis results were removed. The dataset consisted of only letters and words.

Step 2. Conversion Process: Since the product reviews contain uppercase and lowercase inconsistencies, the conversion process was applied, and all letters in the sentences were converted to lowercase. For example, although the words 'satisfied', 'Satisfied', or 'SATISFIED' have a positive meaning, the system can perceive them as different words due to letter differences. For this reason, all sentences were standardized in a uniform form.

Step 3. Removal of Stop-words: It is the process of removing words from the text that do not help to determine the sentiment of the text during the analysis. Stop words are words that do not carry any information and are language-specific functional words that vary according to the language of the text [60,61]. Stop words are words that are programmed to ignore entries both when indexing and when removing them [62]. They lack significant centrality as they return large amounts of redundant information and have no analytical value [23,63].

In this step of the data preprocessing phase, Turkish words frequently used in the comments but not important in terms of sentiment or meaning were removed. Stop-words that did not contribute to the analysis and would cause misleading results were identified as follows and removed from the analysis:

['bir', 've', 'veya', 'de', 'da', 'falan', 'filan', 'felan', 'fln', 'az olsa', 'azda olsa', 'lakin', 'nebze olsa', 'nebzede olsa', 'gerçekten', 'öncelikle', 'gibi', 'gbi', 'çok', 'cok', 'bir', 'bi', 'bır', 'bı', 'vs', 'bence', 'hiç', 'hic', 'hiçbir', 'bu', 'ile', 'ıle', 'ama', 'fakat', 'çünkü', 'eğer', 've ya', 'falanda']

Step 4. Normalization: One of the most important steps in the data preprocessing phase of the study is spelling normalization, which corrects spelling mistakes. Product reviews contain spelling mistakes since they consist of texts produced subjectively by consumers who have not mastered specific grammar rules or are writing randomly. On the other hand, people's use of prolongations in order to express their sentiments in texts also causes spelling mistakes. In the current dataset, the encountered and possible misspellings were corrected and recorded in the system, and the texts were brought to the correct form.

Another reason for the normalization process is the structural form of Turkish. The fact that Turkish is an agglutinative language makes it possible to negate positive words by combining them with different phrases, thus leading to misleading results during the analysis. For example, while 'harika (great)' is a positive word, the sentences 'Harika bir ürün olduğunu söyleyemeyeceğim (I do not think it is a great product)' or 'Harika kapattığını söyleyemeyeceğim (I cannot say that it covered perfectly)' carry negative sentiments. At this stage, 10,040 words and phrases were made suitable for analysis by normalization. Therefore, normalization is an important step of the study. Examples of normalized words and phrases are presented in Table 3.

After the preprocessing, context-specific words and phrases were taken into account while sentiment labeling was performed, and it was aimed to increase the accuracy of the analysis. Since the cosmetics industry is a niche area, it is predicted that accurate results will be obtained with the sentiment dictionary prepared for special purposes. For example, words like yoğun (intense) or parlama (shine) can express a sentiment-laden feature in cosmetic product reviews. While the word 'yoğun' has a positive meaning for an eye make-up product, it may have a negative meaning for a skin care product. The word

'parlama' may have a positive meaning for a body care product but a negative meaning for a skin care product. Thus, it tried to capture all the nuances in the product reviews with manual labeling and custom dictionary creation.

Table 3. Word and phrase normalization list.

Word/Phrase	Normalized Word/Phrase
güzl	güzel (beautiful)
begen	beğen (like)
hizli	hızlı (fast)
ba yıl dım	bayıldım (I loved)
hacimli bir ürün değil	hacim vermedi (didn't give volume)
başarılı bir ürün diyemem	başarısız ürün (unsuccessful product)
surekli	sürekli (consistently)
kullandim	kullandım (used it)
puruzsuz	pürüzsüz (smooth)

At the same time, the unique terminology of cosmetic products and consumer expectations could be considered in the context of jargon originating from specific expressions frequently used in consumer evaluations of cosmetic products. Consumers may focus on specific words and express opinions with detailed information about cosmetic products. For example, in terms of detailed consideration of specific expressions such as 'kapatıcılık' (concealment), 'sabitlenme' (sticking), 'kalıcılık' (persistence), 'portakal kabuğu' (orange peel), the creation of a unique sentiment dictionary with manual labeling enabled the inclusion of meaningful expressions in the model. Thus, a sentiment dictionary consisting of 65,378 words at the word and phrase level was created by manual labeling.

The cosmetic product reviews dataset in each category and product group was divided into 80% training data and 20% test data to be subjected to positive, neutral, and negative classification. Then, the TF-IDF vectorizer method was used to convert the textual data into numerical form and make it suitable for machine learning algorithms. By calculating TF-IDF ratios, the frequency levels and usage percentages of the words in each dataset were calculated. After all these stages, the classification success of manually labeled data with machine learning models was tested with system evaluation.

The confusion matrix results of the K-NN model are presented in Table 4. When the test results are analyzed, the K-NN algorithm's prediction success is high in positive class predictions in each product category. In all product categories, the prediction success of K-NN in the neutral class was lower than in the positive and negative classes. To illustrate, in the lip make-up product category, 10,093 (97.8%) data were correctly predicted in the positive class, and 229 (2.2%) data were incorrectly assigned to other classes. In addition, 69 (14.4%) instances were correctly recognized in the neutral class, 409 (85.6%) were incorrectly predicted in other classes; 549 (31.4%) negative classes were correctly assigned, and 1197 (68.6%) negative data were distributed to other classes. In this case, it is seen that the model fails to learn the neutral class, while the success rate is high in the positive class.

The confusion matrix results showing the success of the SVM model in predicting the classes are given in Table 5. The SVM algorithm achieved very high success in predicting positive classes in each product category. However, it is seen that the model has difficulty in predicting neutral and negative classes. Especially the neutral class was the weak side of the model in all categories. When the success of the SVM model is evaluated with the confusion matrix, it is seen that the model's success in classifying positive examples is relatively high. While 50,829 (98.7%) data were correctly predicted in the positive class, 666 (1.3%) were incorrectly predicted in other classes. In neutral samples, 870 (32.6%) data were predicted in the correct class, while 3133 (67.4%) were in the wrong class, indicating a

low success rate. Finally, 7270 (86%) negative samples were identified in the correct class, while 1182 (14%) samples were mispredicted. This shows that the SVM model does not show added success while working successfully on positive and negative classes.

Table 4. K-NN confusion matrix results for each product group.

Classifier	Product Group	Predicted	Actual		
			Positive	Neutral	Negative
K-Nearest Neighbor	Skin Make-Up	Positive	8541 (95.7%)	256 (2.9%)	128 (1.4%)
		Neutral	264 (63.2%)	106 (25.3%)	48 (11.5%)
		Negative	918 (58.3%)	190 (12.1%)	465 (29.6%)
	Eye Make-Up	Positive	12,476 (97.9%)	132 (1%)	137 (1.1%)
		Neutral	354 (70.6%)	101 (20.2%)	46 (9.2%)
		Negative	1449 (59.8%)	63 (2.6%)	912 (37.6%)
	Lip Make-Up	Positive	10,093 (97.8%)	116 (1.1%)	113 (1.1%)
		Neutral	359 (75.1%)	69 (14.4%)	50 (10.5%)
		Negative	1136 (65.1%)	61 (3.5%)	549 (31.4%)
	Skin Care	Positive	49,878 (96.9%)	781 (1.5%)	836 (1.6%)
		Neutral	1644 (61.6%)	784 (29.4%)	241 (9%)
		Negative	4632 (54.8%)	236 (2.8%)	3584 (42.4%)
	Hair Care	Positive	20,278 (75.8%)	6184 (23.1%)	298 (1.1%)
		Neutral	367 (29.9%)	769 (62.6%)	93 (7.5%)
		Negative	386 (10.4%)	1700 (45.8%)	1626 (43.8%)
	Body Care	Positive	26,622 (83%)	5180 (16%)	321 (1%)
		Neutral	476 (39.6%)	633 (52.6%)	94 (7.8%)
		Negative	418 (14.2%)	1368 (46.8%)	1140 (39%)
	Perfume	Positive	4274 (98.3%)	30 (0.7%)	45 (1%)
		Neutral	123 (82%)	22 (14.7%)	5 (3.3%)
		Negative	588 (69.9%)	10 (1.2%)	243 (28.9%)

Table 5. SVM confusion matrix results for each product group.

Classifier	Product Group	Predicted	Actual		
			Positive	Neutral	Negative
Support Vector Machine	Skin Make-Up	Positive	8816 (98.8%)	4 (0.0%)	105 (1.2%)
		Neutral	290 (69.4%)	61 (14.6%)	67 (16%)
		Negative	295 (18.7%)	6 (0.4%)	1272 (80. 9%)
	Eye Make-Up	Positive	12,577 (98.7%)	10 (0.1%)	158 (1.2%)
		Neutral	338 (67.5%)	88 (17.6%)	75 (14.9%)
		Negative	316 (13%)	3 (0.1%)	2105 (86.9%)
	Lip Make-Up	Positive	10,229 (99.1%)	4 (0.0%)	89 (0.9%)
		Neutral	341 (71.4%)	46 (9.6%)	91 (19%)
		Negative	377(21.5%)	4 (0.2%)	1365 (79.2%)
	Skin Care	Positive	50829 (98.7%)	80 (0.2%)	586 (1.1%)
		Neutral	1334 (50%)	870 (32.6%)	465 (17.4%)
		Negative	1126 (13.3%)	56 (0.7%)	7270 (86%)
	Hair Care	Positive	26,515 (98.1%)	27 (0.1%)	218 (0.8%)
		Neutral	686 (55.8%)	414 (33.7%)	129 (10.5%)
		Negative	603 (16.2%)	21 (0.6%)	3088 (83.2%)
	Body Care	Positive	31,914 (99.3%)	36 (0.1%)	173 (0.5%)
		Neutral	749 (62.2%)	305 (25.4%)	149 (12.4%)
		Negative	534 (18.2%)	11 (0.4%)	2381 (81.4%)
	Perfume	Positive	4279 (98.4%)	0	70 (1.4%)
		Neutral	104 (69.3%)	14 (9.3%)	32 (21.3%)
		Negative	157 (18.7%)	0	684 (81.3%)

The confusion matrix results evaluating the classification prediction success of the decision tree algorithm are presented in Table 6. The DT algorithm successfully predicts positive classes, as seen in the classification performance of other machine learning approaches. Although the positive classification prediction success was high in all groups, confusion in the model was observed in neutral and negative classes. For instance, it was observed that the model's success in recognizing the positive class was high in the skin care products group. While 49,067 (95.3%) data were correctly identified in the positive class, 2428 (4.7%) data were predicted in other classes. In the skin care products group, it was observed that the model's success in recognizing the positive class was high. In the neutral class, 1010 (37.8%) data were correctly predicted, but 1659 (62.2%) data were incorrectly assigned to other classes.

Table 6. DT confusion matrix results for each product group.

Classifier	Product Group	Predicted	Actual		
			Positive	Neutral	Negative
Decision Tree	Skin Make-Up	Positive	8372 (93.8%)	159 (1.8%)	394 (4.4%)
		Neutral	220 (52.6%)	119 (28.4%)	79 (19%)
		Negative	376 (23.9%)	63 (4%)	1134 (72.1%)
	Eye Make-Up	Positive	12,126 (95.1%)	146 (1.2%)	473 (3.7%)
		Neutral	252 (50.3%)	140 (27.9%)	109 (21.8%)
		Negative	376 (15.5%)	52 (2.1%)	1996 (82.4%)
	Lip Make-Up	Positive	9794 (94.9%)	167 (1.6%)	361 (3.5%)
		Neutral	240 (50.2%)	129 (27%)	109 (22.8%)
		Negative	380 (21.8%)	63 (3.6%)	1303 (74.6%)
	Skin Care	Positive	49,067 (95.3%)	864 (1.7%)	1564 (3%)
		Neutral	1157 (43.4%)	1010 (37.8%)	502 (18.8%)
		Negative	1621 (19.2%)	355 (4.2%)	6476 (76.6%)
	Hair Care	Positive	25,672 (95.9%)	378 (1.4%)	710 (2.7%)
		Neutral	543 (44.2%)	489 (39.8%)	197 (16%)
		Negative	729 (19.6%)	128 (3.5%)	2855 (76.9%)
	Body Care	Positive	31,034 (96.6%)	456 (1.4%)	633 (2%)
		Neutral	611 (50.7%)	424 (35.3%)	168 (14%)
		Negative	640 (21.8%)	133 (4.6%)	2153 (73.6%)
	Perfume	Positive	4079 (93.8%)	63 (1.4%)	207 (4.8%)
		Neutral	84 (56%)	34 (22.7%)	32 (21.3%)
		Negative	195 (23.1%)	24 (3%)	622 (73.9%)

When we look at the prediction success of negative product reviews with DT, 2855 data were correctly predicted (76.9%) and 847 (25.1%) reviews were incorrectly predicted. The model successfully predicts positive and negative product reviews in the body care products group. A total of 31,034 (96.6%) reviews were correctly predicted in the positive class, 1089 (3.4%) were distributed to other classes. However, when considered as a ratio, the prediction success of the positive class is high. While 424 (35.3%) neutral examples were predicted correctly, 779 (64.7%) comments were incorrectly assigned to other classes. When we look at the success of the negative class, 2153 (73.6%) comments were predicted in the correct class, and 773 (26.4%) comments were predicted in the wrong classes.

The confusion matrix results of the random forest algorithm in Table 7 conclude that the classification success is high for positive and negative product reviews. However, the classification success is low for neutral reviews. For instance, in the eye make-up products group, it is seen that the RF model again made successful predictions with 12,471 (97.8%) positive comments assigned to the correct class. On the other hand, 274 (2.2%) positive comments were incorrectly assigned to other classes. Although the model predicted 100

(20%) neutral product reviews in the correct class, it incorrectly predicted 401 (80%) product reviews in other classes. On the other hand, while the model correctly predicted 2061 (85%) negative data, 363 (15%) were predicted in the wrong classes, 356 of which were in the positive class.

Table 7. RF confusion matrix results for each product group.

Classifier	Product Group	Predicted	Actual		
			Positive	Neutral	Negative
Random Forest	Skin Make-Up	Positive	8756 (98.1%)	24 (0.3%)	145 (1.6%)
		Neutral	261 (62.4%)	95 (22.7%)	62 (14.9%)
		Negative	416 (26.4%)	7 (0.5%)	1150 (73.1%)
	Eye Make-Up	Positive	12,471 (97.8%)	35 (0.3%)	239 (1.9%)
		Neutral	317 (63.3%)	100 (20%)	84 (16.7%)
		Negative	356 (14.7%)	7 (0.3%)	2061 (85%)
	Lip Make-Up	Positive	10,173 (98.6%)	20 (0.2%)	129 (1.2%)
		Neutral	317 (66.4%)	96 (20%)	65 (13.6%)
		Negative	451 (25.8%)	6 (0.3%)	1289 (73.9%)
	Skin Care	Positive	50,615 (98.3%)	160 (0.3%)	720 (1.4%)
		Neutral	1412 (52.9%)	850 (31.8%)	407 (15.3%)
		Negative	1673 (19.8%)	47 (0.6%)	6732 (79.6%)
	Hair Care	Positive	26,388 (98.6%)	62 (0.2%)	305 (1.2%)
		Neutral	646 (52.6%)	432 (35.1%)	151 (12.3%)
		Negative	724 (19.5%)	34 (0.9%)	2954 (79.6%)
	Body Care	Positive	31,819 (99%)	81 (0.3%)	223 (0.7%)
		Neutral	717 (59.6%)	343 (28.5%)	143 (11.9%)
		Negative	668 (22.8%)	19 (0.7%)	2239 (76.5%)
	Perfume	Positive	4258 (97.9%)	11 (0.3%)	80 (1.8%)
		Neutral	90 (60%)	25 (16.7%)	35 (23.3%)
		Negative	216 (25.7%)	1 (0.1%)	624 (74.2%)

When we look at another category, make-up products, the RF model correctly classified 4258 (97.9%) positive comments and predicted only 91 (2.1%) comments in other classes. Although 25 (16.7%) neutral comments were predicted correctly, 125 (83.3%) data were misclassified. Here, the model also showed low success for neutral comments. For the negative class, 624 (74.2%) data were predicted correctly; however, 216 (25.8%) data were assigned to the positive class even though they were negative, and one negative datum was perceived as neutral.

Table 8 shows the confusion matrix results showing the classification success of the LR model, the last algorithm used in the study. As with the performance of all other algorithms, the logistic regression model showed robust performance in predicting positive comments; however, its performance decreased in the neutral and negative classes.

For example, the LR model successfully predicts positive and negative classes in the skin make-up category. The model predicted 8801 (98.6%) positive product reviews in the correct class and classified 124 (1.4%) reviews in other classes. For the neutral class, 63 (15%) reviews were predicted in the correct class, while 355 (85%) data were distributed into positive and negative classes. Regarding the negative classification success, 1279 (81.3%) data were predicted in the correct class, while 294 (8.7%) were predicted in the wrong classes.

It is seen that the success of predicting the positive and negative class for the eye make-up category is also high. While 12,549 (98.4%) positive comments were correctly classified, 179 (1.4%) data were incorrectly predicted as negative, and 17 (0.2%) data were incorrectly predicted as neutral. While 99 (19.8%) neutral data were predicted in the correct class,

310 (61.9%) neutral data were predicted as positive, and 92 (18.3%) data were incorrectly predicted as negative. A total of 1380 (79%) negative data were correctly predicted; however, 366 (21%) were misclassified, 349 (20%) of which were positive.

Table 8. LR confusion matrix results for each product group.

Classifier	Product Group	Predicted	Actual		
			Positive	Neutral	Negative
Logistic Regression	Skin Make-Up	Positive	8801 (98.6%)	13 (0.2%)	111 (1.2%)
		Neutral	296 (70.8%)	63 (15%)	59 (14.2%)
		Negative	284 (18%)	10 (0.7%)	1279 (81.3%)
	Eye Make-Up	Positive	12,549 (98.4%)	17 (0.2%)	179 (1.4%)
		Neutral	310 (61.9%)	99 (19.8%)	92 (18.3%)
		Negative	298 (12.3%)	21 (0.8%)	2105 (86.9%)
	Lip Make-Up	Positive	10,202 (98.9%)	3 (0.0%)	117 (1.1%)
		Neutral	333 (69.7%)	47 (9.8%)	98 (20.5%)
		Negative	349 (20%)	17 (1%)	1380 (79%)
	Skin Care	Positive	50,615 (98.3%)	160 (0.3%)	720 (1.4%)
		Neutral	1412 (52.9%)	850 (31.8%)	407 (15.3%)
		Negative	1673 (19.8%)	47 (0.6%)	6732 (79.6%)
	Hair Care	Positive	26,434 (98.8%)	47 (0.2%)	279 (1%)
		Neutral	651 (53%)	427 (34.7%)	151 (12.3%)
		Negative	581 (51.6%)	52 (1.4%)	3079 (83%)
	Body Care	Positive	31,864 (99.2%)	52 (0.2%)	207 (0.6%)
		Neutral	720 (59.9%)	340 (28.3%)	143 (11.8%)
		Negative	527 (18%)	52 (1.8%)	2347 (80.2%)
	Perfume	Positive	4285 (98.5%)	1 (0.00%)	63 (1.5%)
		Neutral	105 (70%)	13 (8.7%)	32 (21.3%)
		Negative	147 (17.5%)	4 (0.5%)	690 (82%)

In the last stage of the study, Table 9 presents the classification performance of each machine learning model with accuracy, precision, recall, and f-1 score metrics and a performance comparison of the algorithms.

When all categories are analyzed, each classifier has acceptable and high-classification performance values. However, when compared, the support vector machine has shown high classification performance in all product groups. Although decision tree, random forest, and logistic regression have successful performance measurements, they are relatively behind the support vector machine. On the other hand, K-nearest neighbor did not perform as well as the other classifiers and gave the lowest results in the comparison.

Among machine learning approaches, the algorithms that performed the highest and lowest in all product groups are shown in Table 10. When accuracy rates are examined, the highest score is obtained by SVM, and K-NN obtains the lowest score. On the other hand, in the eye make-up, lip make-up, general make-up, skin care, body care, and perfumery groups, SVM has the highest score. Additionally, with LR and RF in body care and LR in perfumery.

Despite its ability to achieve maximum prediction accuracy when working with extensive and complex data and its tendency to perform better than other algorithms [64], the SVM algorithm has been an important reason for the method to give the best result in the dataset. Table 10 shows that the SVM algorithm has the most accuracy in all product groups and superior performance in other metrics. SVM has come to the forefront by optimizing class boundaries and giving high accuracy rates. At the same time, SVM can obtain true positives and minimize false negatives by having precision, recall, and f-1 score values in all product groups.

Table 9. Proposed model performance metric results of classifiers.

Product Group	Algorithms	Accuracy	Precision	Recall	F-1 Score
Skin Make-Up	K-NN	0.83	0.83	0.83	0.82
	SVM	0.93	0.93	0.93	0.92
	DT	0.88	0.88	0.88	0.88
	RF	0.92	0.91	0.92	0.91
	LR	0.93	0.92	0.93	0.92
Eye Make-Up	K-NN	0.86	0.85	0.86	0.84
	SVM	0.94	0.94	0.94	0.93
	DT	0.91	0.91	0.91	0.91
	RF	0.93	0.93	0.93	0.93
	LR	0.94	0.94	0.94	0.93
Lip Make-Up	K-NN	0.85	0.83	0.85	0.83
	SVM	0.93	0.92	0.93	0.91
	DT	0.89	0.89	0.89	0.89
	RF	0.92	0.92	0.92	0.91
	LR	0.93	0.92	0.93	0.91
Skin Care	K-NN	0.87	0.85	0.87	0.82
	SVM	0.94	0.94	0.94	0.85
	DT	0.90	0.90	0.90	0.94
	RF	0.93	0.93	0.93	0.90
	LR	0.94	0.93	0.94	0.92
Hair Care	K-NN	0.72	0.91	0.72	0.93
	SVM	0.95	0.95	0.95	0.79
	DT	0.92	0.91	0.92	0.94
	RF	0.94	0.94	0.94	0.91
	LR	0.94	0.94	0.94	0.93
Body Care	K-NN	0.78	0.92	0.78	0.94
	SVM	0.95	0.95	0.95	0.84
	DT	0.93	0.92	0.93	0.95
	RF	0.95	0.94	0.95	0.93
	LR	0.95	0.95	0.95	0.94
Perfumery	K-NN	0.85	0.84	0.85	0.95
	SVM	0.93	0.93	0.93	0.82
	DT	0.89	0.88	0.89	0.92
	RF	0.92	0.91	0.92	0.89
	LR	0.93	0.93	0.93	0.91

Table 10. Proposed model performance comparison of classifiers.

Product Group	Accuracy		Precision		Recall		F-1 Score	
	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest
Skin Make-Up	SVM (0.93)	K-NN (0.83)	SVM (0.93)	K-NN (0.83)	SVM, LR (0.93)	K-NN (0.83)	SVM, LR (0.92)	K-NN (0.82)
Eye Make-Up	SVM, LR (0.94)	K-NN (0.86)	SVM, LR (0.94)	K-NN (0.85)	SVM, LR (0.94)	K-NN (0.88)	SVM, RF, LR (0.93)	K-NN (0.84)
Lip Make-Up	SVM, LR (0.93)	K-NN (0.85)	SVM, RF, LR (0.92)	K-NN (0.83)	SVM, LR (0.94)	K-NN (0.85)	SVM, RF, LR (0.91)	K-NN (0.83)
Skin Care	SVM, LR (0.94)	K-NN (0.87)	SVM (0.95)	K-NN (0.84)	SVM, LR (0.94)	K-NN (0.87)	SVM (0.94)	K-NN (0.85)
Hair Care	SVM (0.95)	K-NN (0.72)	SVM (0.95)	K-NN, LR (0.91)	SVM (0.95)	K-NN (0.72)	SVM (0.94)	K-NN, LR (0.79)
Body Care	SVM, RF, LR (0.95)	K-NN (0.78)	SVM, LR (0.95)	K-NN, DT (0.92)	SVM, RF, LR (0.95)	K-NN (0.78)	SVM, LR (0.95)	K-NN (0.84)
Perfumery	SVM, RF, LR (0.93)	K-NN (0.85)	SVM, LR (0.93)	K-NN (0.84)	SVM, RF, LR (0.93)	K-NN (0.85)	SVM, LR (0.92)	K-NN (0.82)

LR is another algorithm with the highest classification performance. LR is an effective method for solving binary or multiple classification problems and has computational efficiency due to its simple implementation [65]. The computational efficiency success of the LR algorithm also manifested itself in the model used in the study.

Among all algorithms, K-NN gave the lowest scores in classification performance. Although K-NN is an important algorithm for density estimation, its efficiency level is low, and its performance varies according to the data size [66]. The high-dimensional and complex structure of all product groups caused the K-NN algorithm to reveal its weakness. The complexity and variety of the data show that the K-NN algorithm has difficulty learning non-linear boundaries. In this sense, SVM and LR were more successful in classifying big data sentiment labels than the K-NN algorithm.

5. Discussion

This study considers three sentiment classification problems, and data are trained by manual tagging. In the literature, [67] performed sentiment analysis on various product groups using RandomizedSearchCV sklearn libraries and obtained the best classification performance result with SVM, similar to the result of this study. In another study on sentiment analysis of various product reviews in Turkish, [68] compared the results of manual and automatic labeling with various deep learning approaches and obtained the best performance result with manual labeling. In [69], with a dimension-based approach, ZEMBEREK performed sentiment analysis with LTSM and CRF models using a Turkish sentiment dictionary and obtained an f-1 score of 86.75%. In another study [70], a new model was proposed for Turkish reviews of electronic products with manual labeling, and an accuracy of 84.23% was obtained using word2vec and RF. As seen in sentiment analysis studies conducted with Turkish data in the literature, Turkish is a language that requires human intervention due to its grammar structure and rich usage. In this study, a human performs manual labeling, and high performance is achieved by learning with various machine learning methods.

The study can be enriched regarding customer satisfaction or product reviews of cosmetic brands with relatively low market share and negative words or can be developed for use in brand comparison research. Due to the large data size, this study detects the general sentiment of consumers. Dimension-based sentiment analysis can be performed by using smaller data sizes or by reducing the existing dataset, and sentiment-attributed features can be revealed with algorithms such as cluster analysis, cause analysis, and latent Dirichlet allocation.

6. Conclusions

The study focuses on processing and analyzing large amounts of Turkish data with text mining and creating a new sentiment dictionary that includes contexts specific to the cosmetics industry through sentiment analysis.

In this study, which was conducted with 875,445 extensive Turkish textual data, the study's first research question (RQ1) can be answered positively by obtaining high performance results. In addition, machine learning approaches were used to evaluate the performance of the sentiment dictionary created from 65,378 words within the scope of RQ2, and it was concluded that the proposed system worked effectively and successfully. Creating a domain-specific Turkish dictionary has significantly contributed to the literature and has been implemented primarily to address the lack of sentiment analysis studies with Turkish texts. This study fills an important gap in the literature, as studies analyzing Turkish texts using untranslated texts have been less researched than analyses in other languages.

Within the scope of RQ3, in the system created with the supervised approach, the data were divided into two, training and test data, and the performance of the system was measured using K-nearest neighbor, support vector machine, decision tree, random forest, and logistic regression algorithms. The algorithms used achieved high accuracy and metric rates in the context of product groups and categories, indicating that the system worked effectively. When the performances of machine learning approaches were evaluated, it was seen that the support vector machine algorithm gave the highest results in product groups. Thus, this study, which significantly contributes to the Turkish sentiment analysis literature, showed that machine learning methods can provide substantial results with Turkish e-commerce data.

Author Contributions: Conceptualization, C.G.Ö. and S.G.; methodology, C.G.Ö.; software, S.G., validation; S.G.; formal analysis, C.G.Ö. and S.G.; investigation, C.G.Ö.; resources, C.G.Ö., data curation, C.G.Ö. and S.G.; writing—original draft preparation, C.G.Ö.; writing—review and editing, S.G. visualization, C.G.Ö.; supervision, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

K-NN	K-Nearest Neighbor
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
LR	Logistic Regression

References

1. Nasreen Taj, M.B.; Girisha, G.S. Insights of strength and weakness of evolving methodologies of sentiment analysis. *Glob. Trans. Proc.* **2021**, *2*, 157–162. [CrossRef]
2. Chachra, A.; Mehndiratta, P.; Gupta, M. Sentiment Analysis of Text Using Deep Convolution Neural Networks. In Proceedings of the 2017 Tenth International Conference on Contemporary Computing (IC3), Nodia, India, 10–12 August 2017.
3. ChandraKala, S.; Sindhu, C. Opinion mining and sentiment classification: A survey. *ICTACT J. Soft Comput.* **2012**, *3*, 420–427.
4. Cambria, E.T. Affective computing and sentiment analysis. *IEEE Int. Syst.* **2016**, *31*, 102. [CrossRef]
5. Singh, Y.; Bhatia, P.K.; Sangwan, O.T. A review of studies on machine learning techniques. *Int. J. Comput. Sci. Secur.* **2007**, *1*, 70–84.
6. Dhage, S.N.; Raina, C.K. A review on machine learning techniques. *Int. J. Recent Innov. Trends Comput. Commun.* **2016**, *4*, 395–399.
7. Tran, Q.-L.; Le, P.T.D.; Do, T.-D. Aspect-based Sentiment Analysis for Vietnamese Reviews about Beauty Product on E-commerce Websites. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, Manila, Philippines, 20–22 October 2020.
8. Salsabila, I.; Sibaroni, Y. Multi aspect sentiment of beauty product reviews using svm and semantic similarity. *RESTI J.* **2021**, *5*, 520–526. [CrossRef]
9. Guen, K.S.; Yujoung, K. Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Inf. Proc. Manag.* **2018**, *54*, 938–957.
10. Clara, A.Y.; Adiwijaya, A.; Purbolaksono, M.D. Aspect based sentiment analysis on beauty product review using random forest. *J. Data Sci. Appl.* **2020**, *3*, 67–77.

11. Fadly; Marlina, D.; Kurniqawan, T.B.; Zakaria, M.Z.; Abdullah, S.F.B. Sentiment analysis on natural skincare products. *J. Data Sci.* **2022**, *12*, 1–17.
12. Park, J. Framework for sentiment-driven evaluation of customer satisfaction with cosmetics brands. *IEEE Access* **2020**, *8*, 98526–98538. [CrossRef]
13. Jaeuhun, P.; Ye-Rim, K.; Su-Bin, K. Customer satisfaction analysis for global cosmetic brands: Text-mining based online review analysis. *J. Korean Soc. Qual. Manag.* **2021**, *49*, 595–607.
14. Hung, C.; Cao, Y.-X. Sentiment classification of Chinese cosmetic reviews based on integration of collocations and concepts. *Electron. Libr.* **2020**, *38*, 145–169. [CrossRef]
15. Romadhony, A.; Faraby, S.A.; Rismala, R.; Wisesti, U.N.; Arifianto, A. Sentiment analysis on a large Indonesian product review dataset. *J. Inf. Syst. Eng. Bus. Int.* **2024**, *10*, 167–178. [CrossRef]
16. Karayiğit, H.; Acı, Ç.; Akdağlı, A. A review of Turkish sentiment analysis and opinion mining. *Balk. J. Electr. Comput. Eng.* **2018**, *6*, 94–98. [CrossRef]
17. Chauhan, P.; Sharma, N.; Sikka, G. The emergence of social media data and sentiment analysis in election prediction. *J. Ambient Int. Hum. Comput.* **2021**, *12*, 2607–2627. [CrossRef]
18. Daud, A.; Khan, W.; Che, D. Urdu language processing: A survey. *Artif. Int. Rev.* **2017**, *47*, 279–311. [CrossRef]
19. Guatam, G.; Yadav, D. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. In Proceedings of the 2014 Seventh International Conference on Contemporary Computing (IC3), Nodia, India, 7–9 August 2014.
20. Schuller, B.; Mousa, A.E.-D.; Vryniotis, V. Sentiment analysis and opinion mining: On optimal parameters and performances. *Wiley Int. Rev. Data Min. Knowl. Discov.* **2015**, *5*, 255–263. [CrossRef]
21. Venugopalan, M.; Gupta, D. Exploring Sentiment Analysis on Twitter Data. In Proceedings of the 2015 Eighth International Conference on Contemporary Computing (IC3), Nodia, India, 20–22 August 2015.
22. Rao, M.V.; Sindhu, C. Detection of Sarcasm on Amazon Product Reviews Using Machine Learning Algorithms under Sentiment Analysis. In Proceedings of the 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 25–27 March 2021.
23. Saritha, B.; Nayak, J. Aspect based sentiment analysis using naïve bayes and support vector classifiers. *Int. J. Anal. Exp. Modal Anal.* **2019**, *11*, 336–344.
24. Kasri, M.; Birjali, M.; Beni-Hssane, A. A Comparison of Features Extraction Methods for Arabic Sentiment Analysis. In Proceedings of the 4th International Conference on Big Data and Internet of Things, Rabat, Morocco, 23–24 October 2019.
25. Manek, A.S.; Shenoy, P.D.; Mohan, M.C.; Venugopal, K.R. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web* **2017**, *20*, 135–154. [CrossRef]
26. Kranjc, J.; Smailović, J.; Podpečan, V.; Grčar, M.; Žnidaršič, N.; Lavrač, N. Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Inf. Proc. Manag.* **2015**, *51*, 187–203. [CrossRef]
27. Keachaou, Z.; Ammar, M.B.B.; Alimi, A.M. Improving E-Learning with Sentiment Analysis of Users' Opinion. In Proceedings of the 2011 IEEE Global Engineering Education Conference (EDUCON), Amman, Jordan, 4–6 April 2011.
28. Xu, G.; Meng, Y.; Qu, X.; Yu, Z.; Wu, Z. Sentiment analysis of comment texts based on BiLSTM. *IEEE Access* **2019**, *7*, 51522–52532. [CrossRef]
29. Dhar, S.; Pednekar, S.; Borad, K.; Save, A. Sentiment Analysis Using Neural Networks: A New Approach. In Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), Coimbatore, India, 20–21 April 2018.
30. Khan, A.; Baharudin, B.; Lee, L.H.; Khan, K. A review of machine learning algorithms for text-documents classification. *J. Adv. Inf. Technol.* **2010**, *1*, 4–20.
31. Hamed, S.; Ezzat, M.; Hefny, H. A review of sentiment analysis techniques. *Int. J. Comput. Appl.* **2020**, *176*, 20–24. [CrossRef]
32. Sánchez-Holgado, P.; Arcila-Calderón, C. Towards the Study of Sentiment in The Public Opinion of Science in Spanish. In Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality—TEEM'18, Salamanca, Spain, 24–26 October 2018.
33. Dey, L.; Chakraborty, S.; Biswass, A.; Bose, B.; Tiwari, S. Sentiment analysis of review datasets using naive bayes and K-NN classifier. *Inf. Retr.* **2016**, *8*, 54–62. [CrossRef]
34. Lee, L.; Pereira, F. Distributional Similarity Models: Clustering vs. Nearest Neighbors. In Proceedings of the 37th Annual Meeting of the ACL, College Park, MA, USA, 20–26 June 1999.
35. Duwairi, R.M.; Qargaz, I. Arabic sentiment analysis using supervised classification. In Proceedings of the International Conference on Future Internet of Things and Cloud, Barcelona, Spain, 24–27 August 2014.
36. Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Trans. Med.* **2016**, *4*, 1–7. [CrossRef] [PubMed]
37. Zhu, S.; Xu, B.; Zheng, D.; Zhao, T. Chinese microblog sentiment analysis based on semi-supervised learning. In *Semantic Web and Web Science*; Li, J., Qi, G., Zhao, D., Nejdl, W., Zheng, H.-T., Eds.; Springer: New York, NY, USA, 2013; pp. 325–331, ISBN 978-1-4614-6879-0.

38. Alzamzami, F.; Hoda, M.; El Saddık, A. Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation. *IEEE Access* **2020**, *8*, 101840–101858. [CrossRef]
39. Zhao, Y.; Dong, S.; Li, L. Sentiment analysis on news comments based on supervised learning method. *Int. J. Multimed. Ubiquitous Eng.* **2014**, *9*, 333–346. [CrossRef]
40. Sohrabi, M.K.; Karimi, F. A feature selection approach to detect spam in the facebook social network. *Arab. J. Sci. Eng.* **2018**, *43*, 949–958. [CrossRef]
41. Hong, S.; Lee, J.; Lee, J.-H. Competitive Self-Training Technique for Sentiment Analysis in Mass Social Media. In Proceedings of the 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS), Kitakyushu, Japan, 3–6 December 2014.
42. Lee, H.-J.; Shin, H.; Hwang, S.-S.; Cho, S.; MacLachlan, D. Semi-supervised response modeling. *J. Interact. Mark.* **2010**, *24*, 42–54. [CrossRef]
43. Bastı, E.; Kuzey, C.; Dursun, D. Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decis. Support Syst.* **2015**, *73*, 15–27. [CrossRef]
44. Zhang, X.; Gong, W.; Kawamura, Y. Customer behavior pattern discovering with web mining. In *Advanced Web Technologies and Applications. APWeb, Lecture Notes in Computer Science*; Yu, J.X., Lin, X., Lu, H., Zhang, Y., Eds.; Springer: Berlin, Germany, 2014; pp. 844–853.
45. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997; ISBN 0071154671.
46. Han, B. Comparison of different machine learning algorithms in classification. *J. Phys. Conf. Ser.* **2021**, *2037*, 012064. [CrossRef]
47. Balaji, T.K.; Annavarapu, C.S.R.; Bablani, A. Machine learning algorithms for social media analysis: A survey. *Comput. Sci. Rev.* **2021**, *40*, 100395.
48. Bahwari, B. Sentiment analysis using random forest algorithm- online social media. *J. Inf. Technol. Its Util.* **2019**, *2*, 29–33.
49. Prabhat, A.; Kullar, V. Sentiment Classification on Big Data Using Naïve Bayes and Logistic Regression. In Proceedings of the 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 5–7 January 2017.
50. Tyagi, A.; Sharma, N. Sentiment analysis using logistic regression and effective word score heuristic. *Int. J. Eng. Technol.* **2018**, *7*, 20–23. [CrossRef]
51. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intel. Rev.* **2022**, *55*, 15731–15780. [CrossRef]
52. Sun, J.; Wang, G.; Cheng, X.; Fu, Y. Mining affective text to improve social media item recommendation. *Inf. Proc. Manag.* **2015**, *51*, 444–457. [CrossRef]
53. Rathan, M.; Hulipalled, V.R.P.; Murugeshwari, P.; Susmitha, M. Every Post Matters: A survey on Applications of Sentiment Analysis in Social Media. In Proceedings of the 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), Bengaluru, India, 17–19 August 2017.
54. de Fortuny, E.J.; Smedt, T.D.; Martens, D.; Daelemans, W. Evaluating and understanding text-based stock price prediction models. *Inf. Proc. Manag.* **2014**, *50*, 426–441. [CrossRef]
55. Giatsoglou, M.; Vozalis, M.G.; Diamantaras, K.; Vakali, A.; Sarigiannidis, G.; Chatzisavvas, K.C. Sentiment analysis leveraging emotions and word embeddings. *Exp. Syst. Appl.* **2017**, *69*, 214–224. [CrossRef]
56. Duwairi, R.; El-Orfali, M. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *J. Inf. Sci.* **2014**, *40*, 501–513. [CrossRef]
57. Wazery, Y.M.; Mohammed, H.S.; Houssein, E.H. Twitter Sentiment Analysis using Deep Neural Network. In Proceedings of the 2018 14th International Computer Engineering Conference (ICENCO), Cairo, Egypt, 29–30 December 2018.
58. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756.
59. Zin, H.M.; Mustapha, N.; Murad, M.A.A.; Sharef, N.M. The Effects of Pre-Processing Strategies in Sentiment Analysis of Online Movie Reviews. In Proceedings of the 2nd International Conference on Applied Science and Technology 2017 (ICAST'17), Kedah, Malaysia, 3–5 April 2017.
60. Jeyapriya, A.; Selvi, K. Extracting Aspects and Mining Opinions in Product Reviews Using Supervised Learning Algorithm. In Proceedings of the 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India, 26–24 February 2015.
61. Bhonde, R.; Bhagwat, B.; Ingulkar, S.; Pandey, A. Sentiment analysis based on dictionary. *Int. J. Emerg. Eng. Res. Technol.* **2015**, *3*, 51–55.
62. Jariwala, V.P. Optimal feature extraction based machine learning approach for sarcasm type detection in news headlines. *Int. J. Comput. Appl.* **2020**, *117*, 25–29.
63. Kawade, D.R.; Oza, K.S. Sentiment analysis: Machine learning approach. *Int. J. Eng. Technol.* **2017**, *9*, 2183–2186. [CrossRef]
64. Osisanwo, F.Y.; Akinsola, J.E.T.; Awodele, O.; Hinmikaiye, J.O.; Olakanmi, O.; Akinjobi, J. Supervised machine learning algorithms: Classification and comparison. *Int. J. Comput. Trends Technol.* **2017**, *48*, 128–138.

65. Premasudha, B.G.; Rampalli, V. A Comparative Study of Logistic Regression, Support Vector Machines, and LSTM Networks for Sentiment Classification in Academic Reviews. In Proceedings of the 2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC), Davangere, India, 24–25 October 2024.
66. Singh, A.; Thakur, N.; Sharma, A. A Review of Supervised Machine Learning Algorithms. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016.
67. Demircan, M.; Seller, A.; Abut, F.; Akay, M.F. Developing Turkish sentiment analysis models using machine learning and e-commerce data. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 202–207. [CrossRef]
68. Çabuk, M.; Yücalar, F.; Toçoğlu, M.A. Makine öğrenmesi ile e-ticaret ürün yorumlarının otomatik analizi. *Avrupa Bilim Ve Teknoloji Dergisi* **2023**, *52*, 110–121.
69. Karagöz, P.; Kama, B.; Öztürk, M.; Toroslu, I.H.; Cantürk, D. A framework for aspect based sentiment analysis on Turkish informal texts. *J. Intel. Inf. Sys.* **2019**, *53*, 431–541. [CrossRef]
70. Pervan, N.; Keleş, H.Y. Sentiment analysis using a random forest classifier on Turkish web comments. *Commun. Fac. Sci. Univ. Ankara Ser. A2–A3 Phys. Sci. Eng.* **2017**, *59*, 69–79.