Acta Infologica,
ACIN 2025, 9 (1): 1–18

https://doi.org/10.26650/acin.1483488
Submitted: 14.05.2024
Revision Requested: 07.12.2024
Last Revision Received: 24.12.2024
Accepted: 31.12.2024
Published Online 03.02.2025

Acta Infologica

Research Article

Open Access

Multi-Category E-Commerce Insights via Social Media Analysis using Machine Learning and BERT

Meryem Gürbüz [1] & Muhammed Kotan [2]

[1] Sakarya University, Institution of Natural Sciences, Sakarya, Türkiye
[2] Sakarya University, Department of Information Systems Engineering, Sakarya, Türkiye

Abstract The burgeoning prevalence of Internet and social media usage has empowered consumers to effortlessly share their opinions about products and services on social media platforms and websites. Consequently, recent research has focused on using machine learning, text mining, and sentiment analysis techniques to extract valuable insights. These insights can then be employed to support businesses in enhancing customer satisfaction and making informed operational and strategic decisions. In this study, a dataset of 5806 Trendyol user reviews was collected from X using the X API within a specified time frame. The dataset was preprocessed and categorized into five predefined categories: product, support, logistics, advertising, and off-topic. Subsequently, the test set was classified using eight machine learning techniques and compared. Finally, sentiment analysis was performed using the pretrained BERTurk model to evaluate user satisfaction and dissatisfaction levels. By integrating machine learning and BERT, this study extracted a general assessment profile of social media users, particularly for e-commerce platforms, and examined social media perspectives on a multi-category basis.

Keywords Sentiment Analysis · Machine Learning · Transformers · User Reviews · E-commerce

Citation: Gürbüz, M. & Kotan, M. (2025). Multi-category e-commerce insights via social media analysis using machine learning and BERT. *Acta Infologica*, *9*(1), 1-18. https://doi.org/10.26650/acin.1483488

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Gürbüz, M. & Kotan, M.

Corresponding author: Muhammed Kotan mkotan@sakarya.edu.tr

Introduction

With the increasing use of the internet and social media, users have started to prefer communication methods established on these platforms rather than traditional ones. Nowadays, people can constantly share their opinions and thoughts on various topics through social media platforms in text, image, or audio format. In this context, X (formerly known as Twitter) allows users to share long textual tweets for longer posts with a character limit of 25,000 characters (X Help Center, 2024). Thanks to this feature, X generates massive datasets, and it is subject to extensive data analysis studies. These vast data volumes have become a significant resource opportunity for various sectors, such as health, education, shopping, production, services, tourism, food, media, finance, and textiles. With the increase in online shopping, people's sharing of opinions, comments, support, or complaints about the products or services they have purchased through social platforms has also increased.

Consumers interact for various purposes, such as sharing their opinions and emotions, providing feedback or complaints about products and services, seeking information, requesting support for after-sales issues, and resolving problems related to their purchases. These interactions produce valuable data that businesses analyze to assess customer satisfaction and to inform strategic and operational decision-making.

Classification of text and reviews in the context of e-commerce is of significant importance for businesses seeking to gain a comprehensive understanding of customer satisfaction and enhance the quality of their offerings. The collection of customer reviews and ratings is a valuable means of obtaining invaluable insights. To effectively process data, various machine learning techniques are frequently used for this purpose in the literature. This study uses eight machine learning techniques to classify social media users' comments about Trendyol into five predefined categories, and the results are compared.

In addition, sentiment analysis plays a significant role in analyzing customers' attitudes and requirements from their reviews. It allows businesses to extract insights from user-generated content like product reviews. The main objective of this endeavor is to obtain and integrate personal viewpoints, emotions and subjective elements from written content, thereby enabling the categorization of content into positive, negative or neutral classifications (Grandi et al., 2015). Businesses prioritize researching and interpreting consumer behaviors to sustain their existence. Therefore, they value customer feedback, opinions, and suggestions on online platforms. Sentiment analysis is an effective method and opportunity for businesses to understand and enhance customer satisfaction, leading to increased competitive advantage and profit margins. This analysis can be executed at different levels, including the document, sentence, and aspect levels, to provide a more detailed understanding of the sentiment (Kotan, 2022). In this study, the Turkish comments of X users were evaluated. The categorical-based sentiment analysis of the comments categorized using machine learning techniques was performed using the Turkish BERT (BERTurk) (Schweter, 2020).

One of the prominent examples of e-commerce platforms in Turkey is Trendyol. Therefore, customers continuously share their opinions, comments, support inquiries, and shipment experiences about Trendyol on various social platforms related to its sales and post-sales processes. These shared posts serve for individuals to communicate directly with each other or with the company. This study analyzes and categorizes tweets to determine the extent to which customers view the offered service policies positively or negatively. The outcomes of this analysis provide strategic insights for decision-making processes.

In contrast to typical sentiment analysis research that applies polarity classification across entire datasets, the proposed method employs a category-based analytical framework. First, we classify user-generated content into a set of predefined categories that reflect key aspects of e-commerce. Then, we perform sentiment analysis at the category level rather than treating all posts as a monolithic dataset. This two-stage approach addresses a common challenge in e-commerce-focused social media analysis, where off-topic or tangential discussions often dilute or misrepresent the overall sentiment. By isolating relevant domains, the proposed method produces precise and sensitive sentiment evaluations. Consequently, stakeholders gain better actionable insights for strategic decision-making because sentiments are linked to meaningful e-commerce categories rather than an aggregated sentiment score spanning disparate or irrelevant comments.

Furthermore, e-commerce platforms can use these results and their analysis to adopt new strategic policies and activities in areas that require improvement, ultimately increasing customer satisfaction and profit margins. Users can freely express their opinions and share their experiences by revealing their identities or remaining anonymous, allowing for a genuine understanding of what users think about the given service, product, or business. Given these reasons, this study holds significance in understanding consumer behaviors and communication preferences for the retail sector.

The remainder of this study will first discuss similar works in the literature. The third section introduces the methodology used in the study. Section 4 discusses the findings. Finally, conclusions and future work are presented in the last section.

Literature Review

Text classification and sentiment analysis are critical tasks in e-commerce, allowing businesses to interpret customer feedback and enhance user satisfaction. In e-commerce, user comments and feedback are analyzed through machine learning algorithms to extract actionable insights that can improve customer experience, marketing strategies, and product development (Qin et al., 2024; George & Baskar, 2024). Traditional machine-learning models are recognized for their interpretability, scalability, and effectiveness in text classification tasks (Shanto et al., 2023).

Traditional machine learning algorithms are particularly well-suited to classifying e-commerce reviews. These models simplify the extraction of structured insights and highlight trends in customer satisfaction and product feedback, assisting companies in identifying key product features that resonate with customers and optimizing product offerings and targeted marketing strategies (Saleem et al., 2019; He et al., 2022). Algorithms such as Random Forest (RF), Naïve Bayes (NB), and Support Vector Machines (SVM) can efficiently process and classify text data when paired with natural language processing techniques like tokenization, stemming, and vectorization. These methods excel with labeled data, making them ideal for binary or multi-class analysis and often require less data to train effectively.

Dharrao et al. (2023) conducted a detailed study on e-commerce product review classification using Amazon reviews, categorizing ratings from one to five. NB, SVM, and Decision Trees (DT) were employed, with DT achieving the highest accuracy and precision. Deniz et al. (2022) created a dataset of over 50,000 Turkish e-commerce reviews, each labeled across multiple categories. The proposed method achieved promising results in terms of identifying nuanced customer opinions by applying multi-label machine learning techniques. They used embedding methods such as Term Frequency–Inverse Document Frequency (TF-IDF), Word2Vec, Global Vectors for Word Representation (GloVe), and Bidirectional Encoder Representations from Transformers (BERT). The machine learning algorithms include RF, SVM, NB, Multi-label k-nearest Neighbors

(Ml-kNN), and various One-vs-Rest (OvsR) classifiers. Polat and Ağca (2022) compared machine learning methods to classify emotional tendencies in TripAdvisor's Turkish and English hotel reviews. Using algorithms such as DT and RF, the analysis revealed that machine-learning approaches outperformed dictionary-based methods in terms of classification accuracy. Yıldız (2020) applied machine learning algorithms, such as RF and Gradient Boosting (GB), to analyze sales effectiveness across product groups in Turkish e-commerce. By examining stock differences based on e-commerce data, this study revealed that the predictive success of algorithms varies depending on the product category, which suggests that factors influencing sales performance may shift across different product groups. Xu et al. (2020) proposed a continuous NB framework for classification that adapts to new domains by using past knowledge and demonstrated robust performance on large-scale multi-domain e-commerce reviews from Amazon products and movies. Kubrusly et al. (2022) applied tree-based methods, such as Classification Tree (CT), RF, GB, and XGBoost (XGB), to a women's clothing e-commerce review dataset. The proposed methods achieved consensus regarding the key classification terms. Chaubey et al. (2023) compared various machine learning techniques for predicting customer purchasing behavior, including supervised classifiers like Logistic Regression (LR), DT, kNN, NB, SVM, RF, Stochastic Gradient Descent (SGD), Artificial Neural Network (ANN), AdaBoost (AB), and XGB, as well as hybrid models like SvmAda, RfAda, and KnnSgd. The hybrid KnnSgd classifier achieved the highest accuracy by combining the kNN and SGD techniques in an ensemble stacking approach. Tang et al. (2015) presented a neural network (NN) approach for predicting review ratings by integrating user-specific information and demonstrated the potential of NNs.

Sentiment analysis can be considered a text classification task involving binary or multi-class categorization and is an important area in e-commerce studies. This enables businesses to interpret customer emotions and opinions expressed in text data, thereby facilitating a better understanding of consumer behavior and enhancing user satisfaction. Huang et al. (2023) conducted a comprehensive study of current sentiment analysis techniques on e-commerce platforms and explored future directions. By analyzing 54 experimental papers selected from 271 papers identified through specific keywords, the research provides valuable insights for researchers. Marong et al. (2020) provided an overview of sentiment analysis and related methodologies for the e-commerce sector. Beyhaqy et al. (2018) analyzed social media users' sentiment toward e-commerce platforms using tweets from Tokopedia and Bukalapak on X. Among the DT, kNN, and NB classifiers, the NB classifier yielded the best results with an accuracy of 77%. Akter et al. (2021) analyzed the sentiment in Bangla text reviews collected from the Bangladeshi e-commerce platform Daraz. The study evaluated several machine learning models, including the RF, LR, SVM, kNN, and XGB models, with the kNN model achieving the highest performance. Demircan et al. (2021) applied SVM, RF, DT, LR, and kNN to classify the sentiments in Turkish product reviews and found that SVM and RF effectively classified the sentiments into three categories. Taşçı et al. (2024) conducted a sentiment analysis of Amazon product reviews using Kaggle data. Applying TF-IDF vectorization and evaluating algorithms like Passive Aggressive (PA), SVM, RF, AB, kNN, and XGB, RF achieved the highest accuracy of 96.13%. Kumar et al. (2016) focused on mining product reviews from websites like Amazon, using algorithms such as NB, LR, and SentiWordNet to classify reviews as positive or negative, with NB proving the most efficient. Shehu and Tokat (2020) examined Turkish tweets by collecting a dataset of 13,000 tweets via the X API. After preprocessing for noise reduction, 10,500 tweets were retained. The hybrid approach combining SVM and RF classifiers achieved the highest accuracy. Rumelli et al. (2019) addressed data imbalance by undersampling and selecting equal numbers of positive and negative samples while discarding neutral reviews. Sentiments were classified using NB, RF, SVM, and kNN. The NB and kNN methods were the fastest, while the SVM and RF methods took longer. Despite the preprocessing challenges and misleading user ratings, NB achieved the highest accuracy. Haque et al. (2018) used Amazon product reviews focusing on electronics, cell phones, accessories, and musical instruments, with a dataset of approximately 48,500 reviews. They applied classifiers, including SVM, NB, SGD, RF, LR, and DT, and they found that linear SVM achieved the best classification performance. Zada and Albayrak (2023) analyzed user comments from a long-standing Turkish online shopping platform. Results were compared with user ratings using RF and AB for classification. Stephenie et al. (2020) used sentiment analysis on Tokopedia product reviews to gauge customer responses to product quality and seller interactions. The optimized RF model achieved a high accuracy of 97.38%. Using a dataset of product reviews from the AliExpress platform, Arobi et al. (2022) employed eight machine learning algorithms to classify the sentiments, and they found that LR was the most effective approach for this dataset. Jabbar et al. (2019) highlighted the use of SVM for constructing a sentiment analysis framework executed on an e-commerce application. This study utilizes online product reviews from Amazon and conducts sentiment analysis at both the review and sentence levels. Şimşek et al. (2023) used Turkish X data to examine users' sentiments on the Hepsiburada e-commerce site. After text mining and evaluation, the tweets were categorized into positive and negative sentiments. Savci and Das (2023) created new Turkish, English, and Arabic datasets, performed comparative sentiment analysis on texts in all three languages, and tested the performance of pretrained models.

The rise of transformer-based language models has dominated NLP studies, becoming a new paradigm in the field (Yildirim & Asgari-Chenaghlu, 2021). Guven (2021a) compared the performance of BERT models against traditional machine learning methods, like RF, NB, and LR, in classifying sentiment in Turkish tweets. BERT achieved a notable accuracy of 98.75%, with LR as the top-performing traditional method at 98.4%, highlighting the efficacy of BERT models for the Turkish language. Koksal et al. (2022) developed a category classification model that uses user comments and product descriptions from an online sales platform. The BERT model trained for the Turkish language provided the best results, forming the basis for a search engine optimization and intent analysis engine specific to the e-commerce sector. Guven (2021b) performed sentiment analysis on product reviews from Hepsiburada to assist consumers in making informed decisions. Initially, RF, NB, and LR were evaluated, with NB achieving the highest accuracy at 89.95%. The study then compared these results with the language models BERT, ELECTRA, and ALBERT and found that ELECTRA outperformed all, with a 92.54% accuracy. Köksal and Özgür (2021) proposed manually annotated Turkish sentiment analysis datasets from X, the BounTi dataset, containing Turkish tweets about specific universities in Turkey. They also evaluated the performance of transformer models, such as MBERT, XLM-Roberta, and BERTurk.

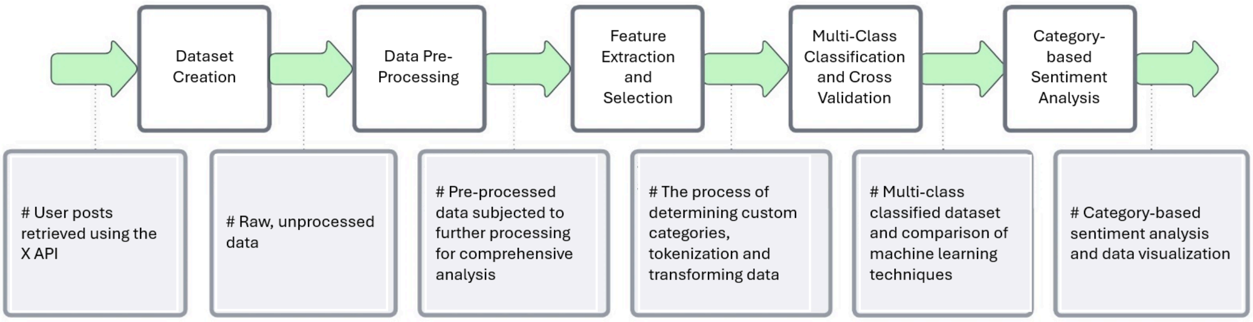
The literature confirms the significance of machine learning techniques in the e-commerce industry, highlighting their practicality and effectiveness in deriving insights that lead to strategic business advantages (Chen, 2022). Traditional machine learning algorithms are widely used due to their high efficiency and interpretability. The development of transformer-based language models represents a paradigm shift in natural language processing. Advanced models, such as BERT and its variants, have demonstrated superior performance in classification tasks across various languages. The proposed method bridges the gap between generalized sentiment analysis and the need for category-specific insights. Many previous studies have focused on sentiment analysis across entire datasets or single-stage classification approaches; however, this study introduces a two-stage methodology to address the limitations of such approaches. In the first stage, user-generated content is classified into predefined categories tailored to e-commerce. In the second stage, sentiment analysis is performed for each category to provide more precise and context-aware evaluations. This dual-stage process tackles the common issue of off-topic or irrelevant content prevalent in social media datasets, which often distorts the overall sentiment analysis. By isolating and analyzing relevant domains, the proposed approach ensures that sentiment evaluations are linked to meaningful e-commerce categories rather than aggregated scores spanning disparate or unrelated comments.

Methodology

In this study, the social media (X) posts of e-commerce customers (Trendyol) were manually labeled with five predefined categories, and the performance of eight machine learning methods for multi-class classification was examined. Subsequently, sentiment analysis was conducted on the categorized comments using the BERT method tailored to the Turkish language. Figure 1 illustrates the key steps of the application process, from data collection and preprocessing to category classification and sentiment analysis. In this study, we used Orange Data Mining software.

Figure 1

Steps of the application process.



Dataset Creation

The vast information sharing on the internet and through social media platforms has led to the accumulation of abundant data on the web. Data extraction and scraping involve retrieving data from websites or other platforms. The dataset used in this study consisted of comments about Trendyol collected through the X API. As a result of this process, 5,806 comments were obtained and stored in .csv format along with their respective dates. Word clouds visually represent text data, which allows researchers to quickly identify the most frequent and important words and phrases. This facilitates exploring datasets and uncovering hidden patterns. The word cloud is presented in Figure 2.

Figure 2

Dataset Wordcloud.

The comments in the dataset were analyzed and manually categorized into five categories. These categories are Advertisement (A), Product (P), Logistics (L), Support (S), and off-topic (O) categories. Descriptions and examples of these categories are presented in Table 1.

Table 1

Descriptions and examples of categories based on human labelling

Category	# of reviews	# of reviews in the test set	Explanation	Example
Advertisement	1243	373	Users' comments on the platform's ads, influencers, and sharing links.	trendyol reklamları başladı!
Product	599	180	Comments about the products	ben siyah istemiştim
Logistics	677	203	Posts about shipping, packaging, and delivery	ay oldu hala ürünüm gelmedi
Support	1159	347	Posts by users and platform team members about support services	siparişimi iptal etseniz?
Off-Topic	2128	638	Related/unrelated comments on specific topics	battım!
Total	5806			

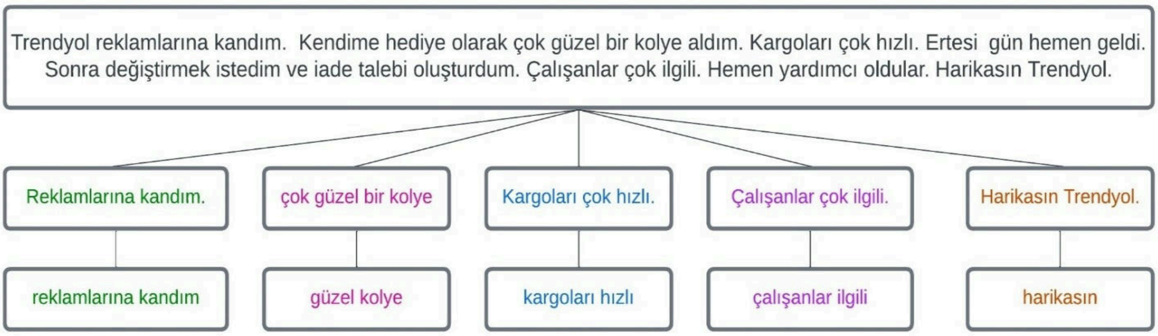
Text Preprocessing

Text preprocessing is the process of cleaning raw data to prepare them for further analysis. These procedures include correction, filling in missing data, removing duplicate data, data transformation, cleaning, normalization, integration, and dimension reduction. Although these processes may vary depending on the specific study, data preprocessing is a crucial step that needs to be carefully executed for more efficient and accurate analysis results. The processes conducted during the data preprocessing stage are illustrated in Figure 3. The following steps were performed during data preprocessing:

- Conversion of uppercase letters to lowercase letters
- Cleaning the punctuation marks and numbers
- Removal of special characters and emojis
- Removal of stop words
- Deletion of empty lines

Figure 3

Sample data preprocessing outcomes.



Feature Extraction and Selection

In this step, TF-IDF was selected. TF-IDF is a widely used technique in text processing and information retrieval to evaluate the importance of a term within a document relative to a collection of documents. The TF-IDF value increases proportionally to the number of times a term appears in a document and is offset by the frequency of the term in the collection. Using the TF-IDF, frequent terms in a document but rare in the overall collection are given higher weights, indicating their importance in that specific document. In contrast, common terms across all documents were penalized in their TF-IDF scores. Figure 4 illustrates the conversion of clear text data to numeric scores using the TF-IDF algorithm.

Figure 4

Example of TF-IDF transformation.

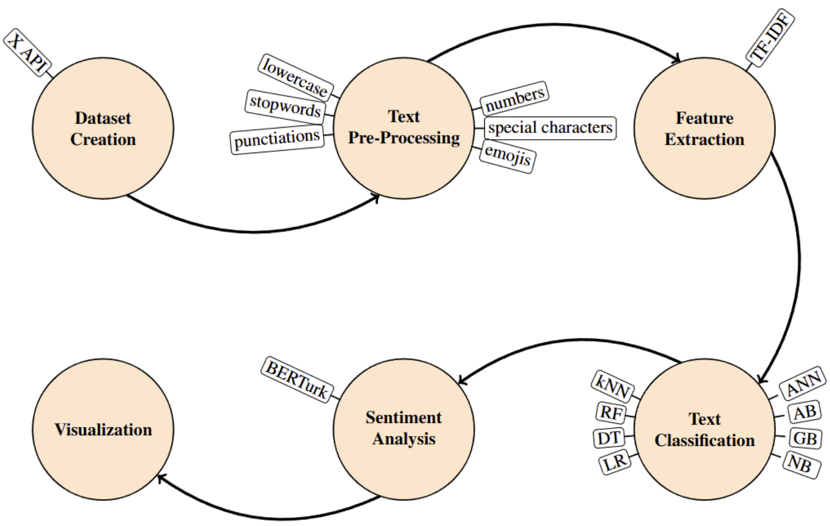
Term	Term ID	TF-IDF Score
kargolamaları	201123	0,56325
iyi	301212	0,89653
satış	125632	0,63254
sonrası	126547	0,31256
destekleri	201123	0,78541
iyi	301212	0,89632
tavsiye	254631	0,79856
ederim	145875	0,35623

Machine-Learning-based Multi-Class Classification

The text classification task assigns predefined categories to the text data. Machine learning algorithms are crucial for automating this process, as they enable computers to analyze large volumes of text and make accurate predictions. The algorithm is presented with a labeled dataset during training, where each text sample is associated with categories. The algorithms learn the patterns and relationships in the data and construct a model that can distinguish different categories. Once trained, the model can be used to predict the category of new, unseen text samples. In this study, AdaBoost (AB), Naïve Bayes (NB), Decision Tree (DT), Artificial Neural Network (ANN), k-Nearest Neighbor (kNN), Logistic Regression (LR), Random Forests (RF) and Gradient Boosting (GB) models were used. The general system structure is summarized in Figure 5.

Figure 5

Overview of the system architecture.



To compare the performance of the methods, a 70%-30% train and test dataset was created. The parametric information of the models and evaluation methods used in this study are presented in Table 2.

Sentiment Analysis using BERTurk

Sentiment analysis is a fundamental aspect of NLP, providing valuable insights for various applications, ranging from analyzing social media content to forecasting stock market trends. Transformer-based BERT (Devlin, 2018) models have demonstrated high accuracy and success in sentiment analysis tasks and have been used in sentiment analysis across various fields, including e-commerce (Sayeed, 2023).

Table 2

Overview of model and parameter settings

MODEL	
k Nearest Neighbor	k = 5; metric: Euclidean
AdaBoost	# of estimators: 50; learning rate: 1.0
Neural Network	# of neurons: 100; max iteration: 200
Decision Tree	Min number of instances in leaves: 2; maximal tree depth: 100
Random Forest	# of trees: 10
Gradient Boosting	# of trees: 100; learning rate: 0.1
Logistic Regression	
Naïve Bayes	
VALIDATION	
Train-Test	%70-%30 (4065 – 1741)

BERT employs different pre-training tasks, such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), to increase the model's understanding of the relationship between words and context. This study used the BERTurk (Schweter, 2020) model and a fine-tuned version (Yildirim, 2024), which is available for Turkish texts. The current version of the model is trained on a filtered and sentence-segmented version of the Turkish OSCAR corpus, a recent Wikipedia dump, various OPUS corpora, and a special corpus (Poyraz, 2022). Yildirim's (2024) fine-tuned BERTurk model is trained with base settings for many downstream tasks and evaluated with a Turkish Benchmark dataset. The study outperformed some baseline approaches for Named-Entity Recognition, Sentiment Analysis, Question Answering, and Text Classification in Turkish Language (Yildirim,2024).

Experimental Results

We comprehensively analyzed e-commerce reviews for Trendyol on X. Machine learning techniques were used to classify user comments by considering predefined classes. Then, a category-based sentiment analysis was conducted. Initially, we employed supervised machine learning algorithms to construct predictive classification models to assess the content of reviews. The manual labeling of the tweets allowed us to train the algorithms effectively. Table 3 summarizes the performance metrics of each model on the multi-classification task.

Table 3

Multi-class classification results

Model	AUC	Accuracy	Precision	Recall	F1 Score
kNN	0.695	0.527	0.702	0.527	0.462
DT	0.835	0.681	0.675	0.681	0.670
RF	0.900	0.729	0.740	0.729	0.720
NN	0.874	0.707	0.715	0.707	0.701
NB	0.907	0.686	0.731	0.686	0.698
LR	0.893	0.731	0.731	0.731	0.726
GB	0.901	0.730	0.760	0.730	0.719
AB	0.897	0.726	0.735	0.726	0.721

The results show that the NB, GB, and RF models achieved the highest performance, with an AUC ≥0.90. The LR model demonstrated competitive performance, closely following with an accuracy of 0.73. The confusion matrices of all models are presented in Table 4. The category-based ROC curves of the LR model are also presented in Figure 6.

The confusion matrices in Table 4 provide a comprehensive view of the models' performances across the predefined categories: S, L, A, O, and P. Each matrix highlights the number of correctly classified instances (diagonal values) and misclassified instances (off-diagonal values), providing insights into the strengths and limitations of each model. The confusion matrices demonstrate that category O consistently achieved the highest correct classification counts across all models, which indicates that its features are relatively distinct. In contrast, category P consistently demonstrated the lowest correct classification counts and was frequently misclassified, which highlights the need for improved feature representation and dataset rebalancing. This analysis provides valuable insights into refining the classification pipeline and optimizing model performance.

Table 4

Confusion matrices of all models

LR	Predicted							GB	Predicted						
	Category	S	L	A	O	P	∑		Category	S	L	A	O	P	∑
Actual	S	278	7	1	51	10	347	Actual	S	267	4	1	68	7	347
	L	31	129	3	31	9	203		L	15	135	0	45	8	203
	A	7	1	279	81	5	373		A	2	0	236	128	7	373
	O	23	17	55	519	24	638		O	20	4	16	589	9	638
	P	15	10	14	73	68	180		P	14	3	5	114	44	180
	∑	354	164	352	755	116	1741		∑	318	146	158	944	75	1741
RF	Predicted							AB	Predicted						
	Category	S	L	A	O	P	∑		Category	S	L	A	O	P	∑
Actual	S	263	8	3	65	8	347	Actual	S	271	8	5	47	16	347
	L	25	131	2	33	12	203		L	32	125	3	31	12	203
	A	3	0	259	102	9	373		A	6	0	248	104	15	373
	O	20	3	26	569	20	638		O	28	6	30	550	24	638
	P	18	6	9	99	48	180		P	21	5	9	75	70	180
	∑	329	148	299	868	97	1741		∑	358	144	295	807	137	1741

NN	Predicted							NB	Predicted						
	Category	S	L	A	O	P	∑		Category	S	L	A	O	P	∑
Actual	S	262	8	2	63	12	347	Actual	S	219	48	4	27	49	347
	L	24	112	6	52	9	203		L	12	153	3	13	22	203
	A	6	0	262	99	6	373		A	2	5	292	41	33	373
	O	25	12	63	525	13	638		O	9	28	77	425	99	638
	P	12	8	15	75	70	180		P	2	16	19	37	106	180
	∑	329	140	348	814	110	1741		∑	244	250	395	543	309	1741
DT	Predicted							kNN	Predicted						
	Category	S	L	A	O	P	∑		Category	S	L	A	O	P	∑
Actual	S	272	11	4	52	8	347	Actual	S	188	1	1	157	0	347
	L	34	121	3	31	14	203		L	8	13	3	178	1	203
	A	13	2	246	104	8	373		A	1	0	79	290	3	373
	O	44	25	49	502	18	638		O	4	0	9	621	4	638
	P	18	24	15	79	44	180		P	2	0	3	159	16	180
	∑	381	183	317	768	92	1741		∑	203	14	95	1405	24	1741

We further conducted sentiment analysis on the classified test set reviews using BERTurk specifically pre-trained on Turkish text to understand the linguistic nuances of the language. The proposed model was employed to predict the sentiment polarity of each review in the test dataset. To evaluate the performance of the sentiment analysis, we compared the predicted sentiments against the ground truth labels in the test dataset. The results indicate that BERTurk achieved an overall accuracy of 77.1%, demonstrating its effectiveness in capturing the sentiment context within Turkish text. Accuracy reflects the model's ability to correctly predict sentiments in most cases. However, it also highlights some challenges, such as language ambiguities and the complexity of specific reviews. The positive-negative ratio of the test set was found to be 34%-66%. Figure 7 presents the sentiment analysis results of the LR-categorized reviews. Figure 7a shows the category-based positive review distributions, and Figure 7b shows the negative distributions. For comparison, the sentiment analysis distributions of the previously classified test dataset based on human-categorized reviews are presented in Figure 8. Figure 8a shows the category-based positive review distributions, and Figure 8b shows the negative distributions.

Figure 6

Category-based ROC curves of the LR model.

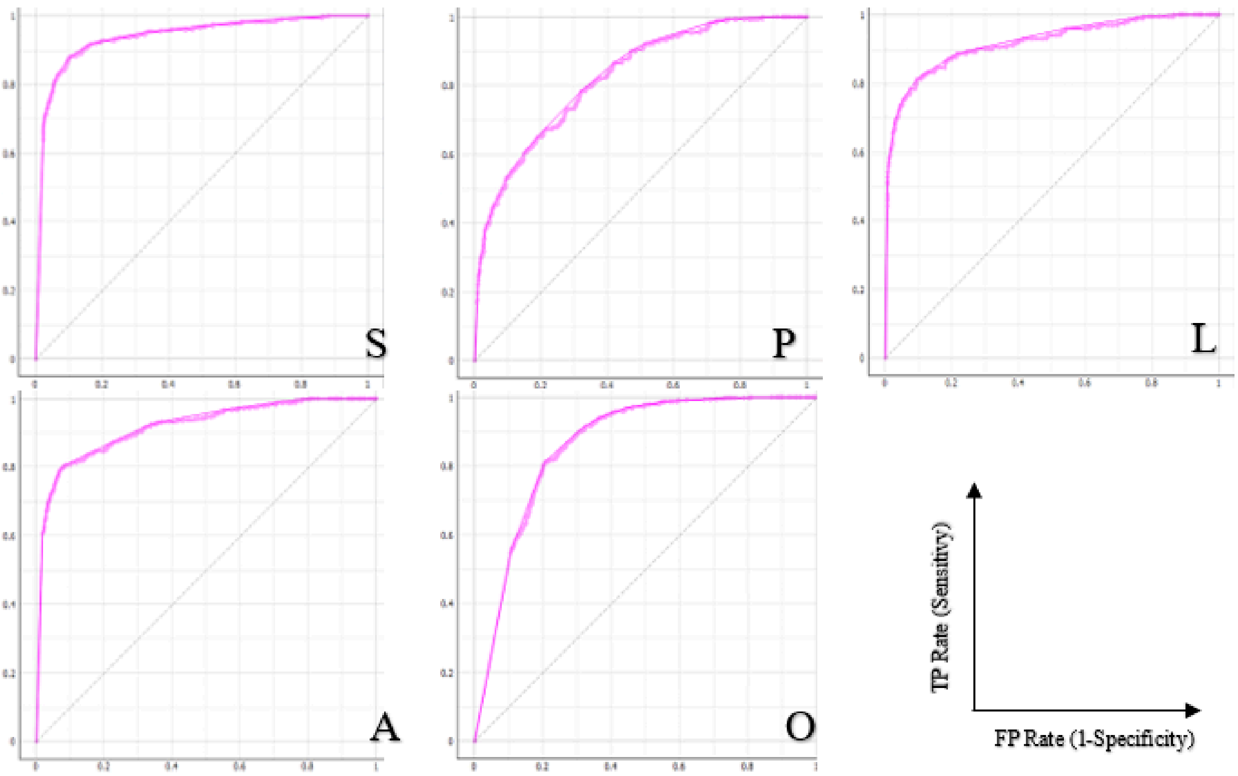
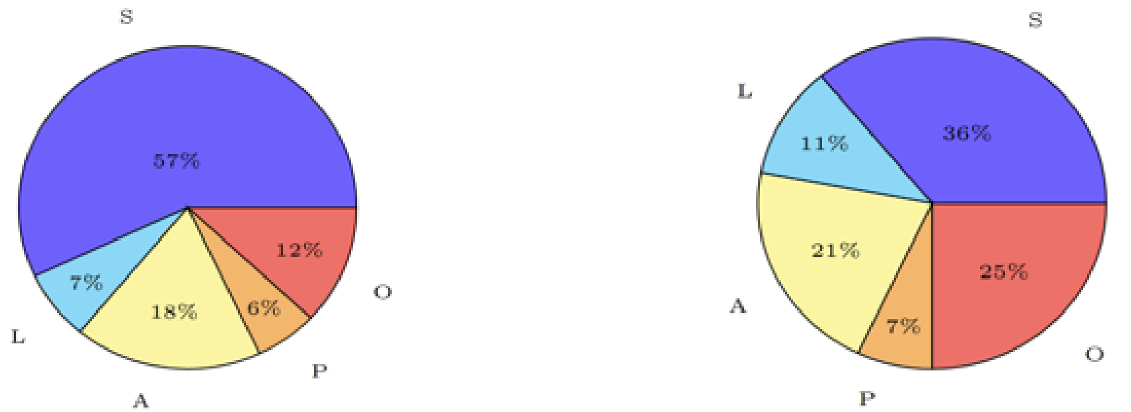


Figure 7

Category-based sentiment distributions of LR-based classification.

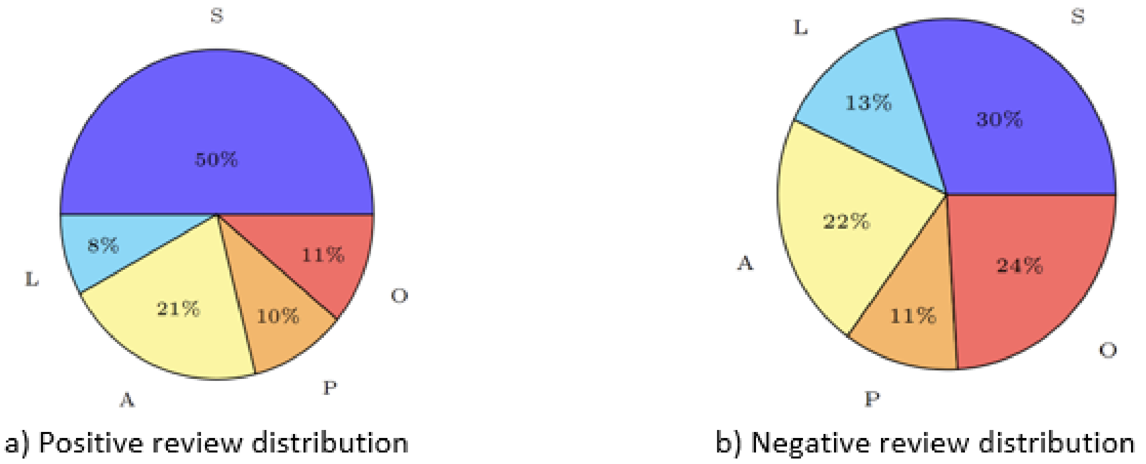


a) Positive review distribution

b) Negative review distribution

Figure 8

Category-based sentiment distributions of human-based classification.



According to the analysis results, comments shared on X about Trendyol have been examined separately, and customer satisfaction has been created for each category. This process attempts to determine the highest and lowest customer satisfaction scores and examine the data. In this context, comments in categories with low customer satisfaction can lead to customer loss. The business can prevent this ongoing loss by implementing improvement and transformation processes within relevant departments, focusing on areas highlighted by negative feedback to increase satisfaction scores.

The results offer several practical benefits for e-commerce platforms. By categorizing user feedback into predefined categories, businesses can better understand the specific aspects of their operations that require improvement. The integration of category-based sentiment analysis also enables businesses to tailor their marketing strategies. By isolating positive sentiment related to specific categories, such as advertising, businesses can identify and replicate successful campaign elements in future promotions. Conversely, identifying dissatisfaction in categories like product or support allows for targeted actions to address customer concerns and enhance brand reputation. Operationally, this approach supports more informed decision-making processes. These insights can also help refine post-sales service strategies, ensuring a seamless customer experience and increasing overall profitability. By offering actionable insights through category-based sentiment analysis, this study bridges the gap between raw data and strategic decision-making. It enables e-commerce platforms to adopt data-driven approaches to enhance their services, customer engagement, and market positioning.

The experimental results demonstrate the effectiveness of machine learning models in classifying e-commerce reviews. In addition, integrating BERTurk for sentiment analysis provided deeper insights into the sentiment distribution in the classified reviews. These findings underscore the potential of combining machine learning algorithms and advanced natural language processing techniques to analyze consumer sentiment on e-commerce platforms.

Discussion and Conclusion

In conclusion, this study investigated the potential of machine learning and sentiment analysis techniques to extract valuable insights from online customer reviews. As businesses increasingly recognize the importance of understanding and responding to customer feedback, the ability to analyze such data effectively has become essential for maintaining competitiveness.

The study successfully employed eight machine learning algorithms for classification tasks using a dataset of 5,806 Trendyol user reviews categorized into five distinct classes. In addition, sentiment analysis using the pretrained BERTurk model provided valuable information about user satisfaction and dissatisfaction levels. This combined approach highlights the effectiveness of integrating machine learning and sentiment analysis to capture a comprehensive assessment profile of social media users about e-commerce platforms. By analyzing user reviews categorized by specific aspects like product, support, advertisement, logistics, and off-topic comments, this study offers a multi-faceted perspective on social media feedback. These insights can help businesses improve customer service, identify areas for improvement, and optimize marketing strategies.

Using traditional machine learning models for multi-category classification and integrating a pretrained BERTurk model for sentiment analysis makes a unique contribution to e-commerce text analysis. Although our dataset from Trendyol presents unique characteristics, such as five distinct thematic categories, the consistency in model performance observed in this study suggests that combining traditional models for category classification with transformer models for sentiment analysis can be a practical framework for similar e-commerce applications. This integrated approach aligns with current research trends and may provide a versatile model structure that can be adapted to other regional or platform-specific datasets, further validating the utility of traditional and transformer-based models in multidimensional text analysis.

The findings demonstrate several practical implications. Categorizing user feedback by specific operational areas allows businesses to pinpoint areas for improvement, such as addressing negative sentiment in logistics or identifying successful advertising strategies. This approach supports strategic decision-making, enabling businesses to optimize resource allocation, refine marketing strategies, and enhance customer satisfaction. By offering a granular perspective on sentiment, this study bridges the gap between raw social media data and actionable business insights, helping e-commerce platforms improve customer engagement and market positioning.

Our study introduces a unique approach by combining traditional machine learning models for category classification with a pretrained BERTurk model for sentiment analysis; however, some limitations must be acknowledged. A larger dataset would provide a broader spectrum of language patterns, thereby enhancing model robustness and generalizability. The traditional machine learning models used for classification were not extensively optimized, suggesting opportunities for further improvement through hyperparameter tuning or ensemble techniques. Our study also relies on different methodologies for category classification and sentiment analysis, with traditional machine learning models and a transformer model. Although this layered approach allows us to use the strengths of both techniques, it may increase the complexity of the classification process. Future research could explore using a single, unified model to handle category and sentiment classification, such as fine-tuning transformer-based models to perform both tasks simultaneously. This approach could streamline the process and potentially yield improved classification accuracy. In addition, although BERTurk performed well on e-commerce-specific data, fine-tuning the model could improve its sensitivity to domain-specific nuances. Addressing these limitations in future studies could enhance our approach's accuracy, adaptability, and scalability, further validating its effectiveness across larger and more diverse datasets.

Future research could focus on integrating classical and advanced machine learning techniques to address data preprocessing and class imbalance challenges and extend the methodology to multilingual datasets. Incorporating advanced models, such as RoBERTa or multilingual BERT variants, may further enhance classification accuracy and the depth of sentiment analysis. In addition, tracking the evolution of user sentiment over time and using fine-grained sentiment analysis techniques can provide richer insights into consumer behavior across diverse markets. Applying this framework to a broader range of e-commerce platforms and product categories could uncover more nuanced feedback patterns, ultimately extending the utility and applicability of this approach to understanding customer preferences and improving e-commerce strategies.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- M.G., M.K.; Data Acquisition- M.G.; Data Analysis/Interpretation- M.G., M.K.; Drafting Manuscript- M.G., M.K.; Critical Revision of Manuscript- M.K.; Final Approval and Accountability- M.G., M.K.; Technical or Material Support- M.G., M.K.; Supervision- M.K.

Conflict of Interest: The authors have no conflict of interest to declare

Grant Support: The authors declared that this study has received no financial support.

Author Details

Meryem Gürbüz

[1] Sakarya University, Institution of Natural Sciences, Sakarya, Türkiye

0000-0003-1354-9951

Muhammed Kotan

[2] Sakarya University, Department of Information Systems Engineering, Sakarya, Türkiye

0000-0002-5218-8848 mkotan@sakarya.edu.tr

References

Arobi, R., Rijon, R. H., Haque, N., Safkat, T. I. M., & Safoan, S. (2022). Sentiment analysis on E-commerce based product reviews using machine learning algorithms (Doctoral dissertation, Brac University).

Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. *In 2014 sixth international conference on communication systems and networks (COMSNETS)* (pp. 1-8). IEEE.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018, October). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. *In 2018 international conference on orange technologies (ICOT)* (pp. 1-6). IEEE.

Chaubey, G., Gavhane, P. R., Bisen, D., & Arjaria, S. K. (2023). Customer purchasing behavior prediction using machine learning classification techniques. *Journal of Ambient Intelligence and Humanized Computing,* 14(12), 16133-16157.

Chen, N. (2022). Research on E-Commerce Database Marketing Based on Machine Learning Algorithm. *Computational Intelligence and Neuroscience,* 2022(1), 7973446.

Demircan, M., Seller, A., Abut, F., & Akay, M. F. (2021). Developing Turkish sentiment analysis models using machine learning and e-commerce data. *International Journal of Cognitive Computing in Engineering, 2,* 202-207.

Deniz, E., Erbay, H., & Coşar, M. (2022). Multi-label classification of e-commerce customer reviews via machine learning. *Axioms, 11*(9), 436.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Dharrao, D., Deokate, S., Bongale, A. M., & Urolagin, S. (2023, March). E-commerce product review classification based on supervised machine learning techniques. *In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1934-1939). IEEE.

George, A. S., & Baskar, T. (2024). Leveraging Big Data and Sentiment Analysis for Actionable Insights: A Review of Data Mining Approaches for Social Media. *Partners Universal International Innovation Journal,* 2(4), 39-59.

Grandi, U., Loreggia, A., Rossi, F., & Saraswat, V. (2015). A borda count for collective sentiment analysis. *Annals of Mathematics and Artificial Intelligence, 77*(3-4), 281-302. https://doi.org/10.1007/s10472-015-9488-0

Gound, R. S., Tikone, P. V., Suryawanshi, S. S., & Nagpal, D. (2018). Twitter Data Sentiment Analysis and Visualization. *International Journal of Computer Applications,* 975, 8887.

Guven, Z. A. (2021a). The effect of bert, electra and albert language models on sentiment analysis for turkish product reviews. *In 2021 6th international conference on computer science and engineering (UBMK)* (pp. 629-632). IEEE.

Guven, Z. A. (2021b). Comparison of BERT models and machine learning methods for sentiment analysis on Turkish tweets. *In 2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 98-101). IEEE.

Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. *In 2018 IEEE international conference on innovative research and development (ICIRD)* (pp. 1-6). IEEE.

He, H., Zhou, G., & Zhao, S. (2022). Exploring e-commerce product experience based on fusion sentiment analysis method. *IEEE Access*, 10, 110248-110260.

Huang, H., Asemi, A., & Mustafa, M. B. (2023). Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. *IEEE Access*.

Jabbar, J., Urooj, I., JunSheng, W., & Azeem, N. (2019). Real-time sentiment analysis on E-commerce application. *In 2019 IEEE 16th international conference on networking, sensing and control (ICNSC)* (pp. 391-396). IEEE.

Koksal, D., Alacan, M. M., Olgun, E., & Sakar, C. O. (2022). Natural Language Processing-Based Product Category Classification Model for E-Commerce. *In 2022 30th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Köksal, A., & Özgür, A. (2021). Twitter dataset and evaluation of transformers for Turkish sentiment analysis. *In 2021 29th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Kotan, M. (2022). Duygu Analizi ve Dijital Dönüşüm Üzerine Etkileri. *In Book: Dijital Etkileşimler ve Sektörel Yansımaları II.* Gazi Kitapevi.

Kubrusly, J., Neves, A. L., & Marques, T. L. (2022). A statistical analysis of textual e-commerce reviews using tree-based methods. *Open Journal of Statistics,* 12(3), 357-372.

Kumar, K. S., Desai, J., & Majumdar, J. (2016, December). Opinion mining and sentiment analysis on online customer review. *In 2016 IEEE international conference on computational intelligence and computing research (ICCIC)* (pp. 1-4). IEEE.

Marong, M., Batcha, N. K., & Mafas, R. (2020). Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms. *Journal of Applied Technology and Innovation,* 4(1), 6.

Polat, H., & Ağca, Y. (2022). Tripadvisor Kullanıcılarının Türkçe Ve İngilizce Yorumları Kapsamında Duygu Analizi Yöntemlerinin Karşılaştırmalı Analizi. *Abant sosyal bilimler dergisi*, 22(2), 901-916.

Poyraz, Y. Z., Tugcu, M., & Amasyali, M. F. (2022). Improving BERT Pre-training with Hard Negative Pairs. *In 2022 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

Rumelli, M., Akkuş, D., Kart, Ö., & Isik, Z. (2019, October). Sentiment analysis in Turkish text with machine learning algorithms. *In 2019 Innovations in intelligent systems and applications conference (ASYU)* (pp. 1-5). IEEE.

Qin, Z., Dong, K., & Xie, B. (2024). What affects customers' online shopping behavior, research that applied machine learning to Amazon product reviews. *Applied and Computational Engineering*, 76, 48-64.

Saleem, H., Muhammad, K. B., Nizamani, A. H., Saleem, S., & Aslam, A. M. (2019). Data science and machine learning approach to improve E-commerce sales performance on social web. *International Journal of Computer Science and Network Security (IJCSNS)*, 19.

Savci, P., & Das, B. (2023). Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. *Journal of King Saud University-Computer and Information Sciences, 35*(3), 227-237.

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A Review of Applications in Sentiment Analysis. *HighTech and Innovation Journal,* 4(2), 453-462.

Schweter, S. (2020). Berturk—Bert models for turkish. *Zenodo.* https://doi.org/10.5281/zenodo.3770924

Shanto, S. S., Ahmed, Z., & Jony, A. I. (2023). Mining User Opinions: A Balanced Bangla Sentiment Analysis Dataset for E-Commerce. *Malaysian Journal of Science and Advanced Technology*, 272-279.

Shehu, H. A., & Tokat, S. (2020). A hybrid approach for the sentiment analysis of Turkish Twitter data. *In Artificial Intelligence and Applied Mathematics in Engineering Problems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2019)* (pp. 182-190). Springer International Publishing.

Şimşek, İ., Kökçam, A. H., Demir, H. I., Erden, C. (2023). Sentiment Analysis of Twitter Data of Hepsiburada E-commerce Site Customers with Natural Language Processing. *In International Symposium on Intelligent Manufacturing and Service Systems* (pp. 567-578). Singapore: Springer Nature Singapore.

Tang, D., Qin, B., Liu, T., & Yang, Y. (2015, June). User modeling with neural network for review rating prediction. *In Twenty-fourth international joint conference on artificial intelligence*.

Taşcı, M. E., Rasheed, J., & Özkul, T. (2024, January). Sentiment Analysis on Reviews of Amazon Products Using Different Machine Learning Algorithms. *In International Conference on Forthcoming Networks and Sustainability in the AIoT Era* (pp. 318-327). Cham: Springer Nature Switzerland.

Stephenie, Warsito, B., & Prahutama, A. (2020). Sentiment analysis on tokopedia product online reviews using random forest method. *In E3S Web of Conferences* (Vol. 202, p. 16006). EDP Sciences.

X Help Center. (2024). About different types of posts. Retrieved from https://help.twitter.com/en/using-x/types-of-posts

Xu, F., Pan, Z., & Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management*, 57(5), 102221.

Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access, 8*, 23522-23530.

Yildırım, S. (2024). Fine-tuning Transformer-based Encoder for Turkish Language Understanding Tasks. *arXiv preprint* arXiv:2401.17396.

Yıldırım, S., & Asgari-Chenaghlu, M. (2021). Mastering Transformers: Build state-of-the-art models from scratch with advanced natural language processing techniques. Packt Publishing Ltd.

Yıldız, H. (2020). The Role Of Machine Learning Algorithms In Determining Product Sales In E-Commerce: A Case Study For Turkey. *Current and Historical Debates in Social Sciences: Field Studies and Analysis*, 175.

Zada, A. J. J., & Albayrak, A. (2023). Duygu Analizi ve Topluluk Öğrenmesi Yaklaşımları ile Kullanıcı Yorumlarının Analizi. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 11(4), 1725-1732.