

# BERT-Based Sentiment Analysis of Turkish e-Commerce Reviews: Star Ratings Versus Text

Ayşe Öcal<sup>1\*</sup> 

<sup>1</sup>Faculty Of Electrical, Electronics, Computer Engineering, Yıldız Technical University, İstanbul, Türkiye, [ror.org/0547yzj13](http://ror.org/0547yzj13)

Corresponding author:

Ayşe Öcal,  
Faculty of Electrical, Electronics,  
Computer Engineering,  
Yıldız Technical University,  
İstanbul, Türkiye  
[ayse.ocal@yildiz.edu.tr](mailto:ayse.ocal@yildiz.edu.tr)



Article History:

Received: 21.07.2025  
Revised: 20.08.2025  
Accepted: 25.08.2025  
Published Online: 13.10.2025

## ABSTRACT

This study examines sentiment analysis in Turkish e-commerce product reviews by comparing two distinct approaches: classification based on star ratings and textual sentiment using a BERT-based model. Two models were fine-tuned for this purpose: Model 1, trained on numerical star ratings, and Model 2, trained on manually labeled sentiment in review texts, to evaluate their performance in accurately capturing customer sentiment. The results reveal that star ratings often fail to reflect true sentiment, as many users assign high ratings despite expressing negative opinions in the text. Model 1 tended to overclassify reviews as negative, while Model 2, which used direct text sentiment labels, provided a more balanced classification across sentiment categories. Chi-square tests confirmed a statistically significant difference between the predictions of the two models, highlighting the impact of labeling methods on model behavior. Furthermore, our findings reinforce the value of deep learning approaches, particularly transformer-based models like BERT, in processing Turkish-language texts, which pose challenges for traditional dictionary-based methods due to their complex morphology and syntax. From a business perspective, relying solely on star ratings may lead to an inaccurate interpretation of sentiment. Incorporating text-based analysis can offer more precise insights into customer satisfaction. Future research may explore multimodal sentiment analysis by integrating visual or video data and examining how AI-driven sentiment systems influence decision-making processes across different sectors.

**Keywords:** Sentiment analysis, BERT, Turkish language, E-Commerce, Natural language processing

## 1. Introduction

In today's digital era, numerous activities take place online, including e-commerce [1]. E-commerce is a digital business strategy that enables online buying and selling of goods and services. With the widespread use of the Internet, e-commerce has become a crucial component of the modern commercial landscape, driven by factors such as the increasing prevalence of mobile devices and applications, competitive pricing, 24/7 availability, the ability to compare products and alternatives, simple return policies, ease of access, and fast transactions [1], [2], [3]. As a result, e-commerce continues to grow in popularity [2], [3].

One of the most common e-commerce business models is the business-to-consumer (B2C) model, where companies sell goods or services directly to customers. In this model, businesses interact directly with consumers, making their products and services easily accessible. This form of e-commerce streamlines the purchasing process, allowing customers to buy products or services online conveniently. Companies operating under the B2C model include Amazon, eBay, Airbnb, Alibaba, Hepsiburada, Netflix, N11, and Trendyol. For these businesses, collecting and analyzing customer feedback is crucial in shaping effective marketing strategies, meeting customer needs, and maintaining a competitive edge [2].

Customer feedback enables businesses to enhance their products and services, ultimately improving customer satisfaction. Although companies use surveys, complaint forms, and social media interactions to gather feedback, online product reviews are the most important sources of information in this context [2]. Customer reviews provide valuable insights into product quality, performance, pricing, and customer service [3], [4]. Moreover, reviews inform businesses and influence other customers' purchasing decisions by providing different perspectives on a product's quality and performance [2], [3]. While positive reviews can boost consumer confidence, negative reviews can deter potential buyers. This phenomenon, explained by the concept of electronic word-of-mouth communication [1], refers to the way potential customers interested in similar products or services share their opinions and experiences through digital platforms [1], [2]. In such an environment, businesses must continuously adapt to the evolving expectations of their customers [1]. Businesses can improve their products and enhance customer satisfaction by incorporating customer feedback into their decision-making processes.

With the rapid expansion of e-commerce platforms, the volume of product reviews has significantly increased, with hundreds of new comments posted every second [5]. Given this sheer volume, manually analyzing these reviews is nearly impossible. At this stage, natural language processing (NLP) techniques, particularly sentiment analysis, enable the automatic extraction of sentiment from large-scale textual data, including product reviews, movie reviews, and online discussions [6], [7], [8], [9], [10].

In e-commerce, sentiment analysis allows companies to extract valuable insights from vast amounts of customer feedback, enabling them to make data-driven improvements to their products and services. Sentiment analysis involves identifying and categorizing expressions to determine a person's attitude toward a particular topic, product, or service—typically classified as positive, negative, or neutral. It is widely used across various domains, including healthcare, entertainment, education, social media, and e-commerce, to analyze public sentiment on social, political, and commercial topics [6], [7], [11]. This helps researchers, businesses, and policymakers understand public opinion and make more informed and responsive decisions in an increasingly digital and interconnected world [6], [7]. In e-commerce, sentiment analysis allows companies to extract valuable insights from vast amounts of customer feedback, enabling them to make data-driven improvements to their products and services.

While sentiment analysis is traditionally text-based, recent advancements have introduced multimodal sentiment analysis, which incorporates multiple data sources, including text, images, audio, and video [6]. However, the most widely used approach remains text-based sentiment analysis, which predicts the sentiment of textual data [6]. Sentiment analysis can be conducted at various levels: document, sentence, phrase, and word. For example, product reviews and social media comments are often treated as short documents, and sentiment analysis is typically performed at the document level in such cases [6], [7], [9]. Several techniques are employed for sentiment analysis, including dictionary-based approaches (e.g., TextBlob, VADER), machine learning models, deep learning models, and hybrid approaches that combine these techniques [6].

Existing literature on sentiment analysis in Turkish texts has primarily employed dictionary-based methods. For example, dictionary-based sentiment analysis has been employed to analyze customer reviews [12], social media data [13], movie reviews, and social media discussions [14] related to the Syrian crisis and refugees [15]. Despite their popularity, dictionary-based methods have notable limitations. One major drawback is their inability to capture context-dependent meanings, as the same word can have different connotations depending on the context [7], [16], [17]. Furthermore, these methods fail to account for word order and syntactic structure, which are crucial for accurately determining sentiment.

In response to these challenges, more advanced sentiment analysis methods have recently been applied to Turkish text data [5]. For instance, experiments were conducted on Turkish movie and hotel reviews using BERT (Bidirectional Encoder Representations from Transformers) models [18], [19], classifying each review as positive or negative and demonstrating the effectiveness of such models. However, relatively few studies have focused on Turkish e-commerce reviews. One notable study [20] compared deep learning architectures, including recurrent neural networks (RNNs) and long short-term memory (LSTM) units, for sentiment analysis on a dataset of Turkish e-commerce reviews. Another study [5] employed supervised machine learning models—including support vector machines (SVM), random forests (RF), decision trees (DT), logistic regression (LR), and k-nearest neighbors (KNN)—to classify product reviews from hepsiburada.com into positive, negative, and neutral categories based on star ratings. Another study [21] analyzed Trendyol product reviews by evaluating sentiment based on the frequency of specific words in the review text. Their findings suggest that star ratings can sometimes be misleading. For instance, a customer may rate a product 5 stars to increase the visibility of their review, despite expressing a negative opinion in the text. This issue poses a significant challenge for businesses seeking reliable customer feedback and remains an area that requires further research.

Thus, in this study, we aimed to investigate whether there is a significant inconsistency between review ratings and text, which could lead to misleading sentiment analysis results. To address this issue, we developed two BERT-based sentiment analysis models to analyze product reviews collected from Trendyol. Given the demonstrated effectiveness of BERT models for sentiment analysis [7], [19], [22], [23], we selected this approach. The first BERT model was fine-tuned using training data where sentiment (positive or negative) was determined based on the star ratings provided by customers. The second BERT model was fine-tuned using training data where sentiment was directly labeled based on the review text. The details of the dataset and methodology are presented in the following sections.

The remainder of this paper is organized as follows: Section 2 describes the datasets and research methodology. Section 3 presents the results. Section 4 discusses key findings, contributions, and future research directions. Finally, Section 5 concludes the paper.

## 2. Materials and Methods

The methodology followed in this research is presented in Figure 1. The research framework consists of four tasks:

- i) generating two datasets to compare the impact of the datasets,
- ii) fine-tuning two BERT models with the datasets generated in the first step,
- iii) evaluating results with performance metrics, and
- iv) Chi-square tests will be conducted to compare the models and explore whether an inconsistency occurs between review ratings and text.

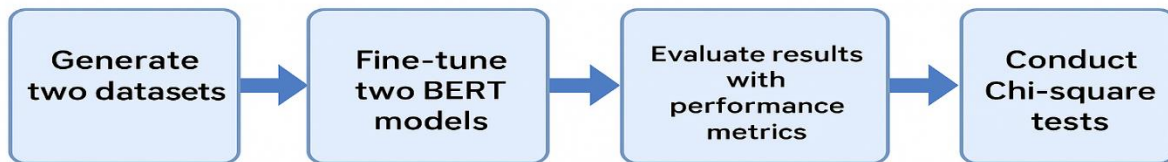


Figure 1. Methodology Workflow

The procedures for data extraction and model fine-tuning are elaborated in the subsequent sections.

## 2.1 Data Extraction

We collected data from the Trendyol website. Trendyol is an e-commerce platform with more than 30 million customers and 80 million products, and it is one of the leading online shopping venues in Türkiye [24], [25]. This platform allows its customers to share their experiences online, resulting in many customer reviews about the products they purchase. To extract data in this study, we used Trendyol's open API.

Our approach adheres to ethical guidelines by fully complying with Trendyol's terms of service and privacy policies. We prioritized user privacy by anonymizing all personally identifiable customer information. Moreover, we consulted Trendyol representatives, who confirmed that this study does not require an ethics committee approval. Transparency is also important for us; therefore, we disclosed the source of the data and clarified the purpose of its use in our analysis. The platform's name is explicitly mentioned to ensure transparency, facilitate the reproducibility of the research, and promote scientific openness. This study is conducted solely as a case study for academic purposes and does not contain any form of advertisement or promotional content.

Additionally, we avoided manipulative practices by presenting the findings unbiasedly, refraining from distorting the reviews. The insights derived from these data aim to contribute to an understanding of consumer behavior and inform market strategies. This commitment to ethical data usage ensures that our research is conducted with integrity, respect for privacy, and a focus on the responsible utilization of publicly available information.

## 2.2 Datasets

### 2.2.1 Dataset for Fine-Tuning the First BERT Model

For the first model, we manually constructed a dataset from the Trendyol e-commerce platform. Two product categories—vacuum cleaners and mobile phones—were selected for analysis. These categories were chosen because they are among the most frequently reviewed technology products on the platform, providing a high volume of customer feedback and ensuring a sufficiently large and balanced dataset for model training. A total of 96 products from these two categories were included.

To construct the dataset, reviews were collected across all five rating levels (1 to 5 stars). We aimed to balance the dataset by collecting approximately equal numbers of reviews for each star rating. The distribution of the collected data was as follows:

- 1 star: 10,215 reviews
- 2 stars: 6,278 reviews
- 3 stars: 10,019 reviews
- 4 stars: 10,345 reviews
- 5 stars: 10,005 reviews

For sentiment classification, reviews with 1 and 2 stars were labeled as negative, while those with 4 and 5 stars were labeled as positive. Reviews with 3 stars were excluded to avoid neutral sentiment ambiguity. The dataset was further cleaned by removing emojis, duplicate entries, and null values. After preprocessing, the final dataset contained 29,473 reviews (13,191 negative and 16,282 positive).

The dataset was split into training (80%) and testing (20%) subsets using the sklearn library. The evaluation results of the first model, presented in the Results section, are based on the held-out test set. Although limited to two categories, the primary purpose of this dataset is not to evaluate specific products. The primary goal is to construct a dataset based on star ratings (1–5), ensuring a clear and balanced sentiment distribution, rather than emphasizing specific categories or individual products.

### 2.2.2 Dataset for Fine-Tuning the Second BERT Model

For the second BERT model, we used a publicly available dataset obtained from GitHub. This dataset comprises 2,000 comments collected from three e-commerce platforms in Türkiye: Hepsiburada, N11, and Trendyol. Customer comments

were labeled as positive or negative based on the review text, with equal samples in both categories; however, the specific product categories or items were not disclosed. Since the main objective of this study is not to analyze product-specific trends but to perform sentiment classification on textual data, the lack of product category information does not affect the validity of the dataset for fine-tuning purposes. What matters here is the presence of sufficient and balanced text samples that enable sentiment analysis rather than the type of products reviewed. The dataset was split into training and test sets using the sklearn library, with 80% allocated for training and 20% for testing. The evaluation performance of the second model is based on the test set.

Table 1. Dataset composition for fine-tuning the second sentiment analysis model

Platform	Positive Samples	Negative Samples
hepsiburada.com	174	109
n11.com	253	705
trendyol.com	573	186
Total	1000	1000

*Note.* This table presents the positive and negative samples collected from each Turkish e-commerce platform for fine-tuning Model 2.

### 2.2.3 Third Dataset Used for Testing Both Models

We produced a third dataset to compare the two fine-tuned models. For this purpose, we randomly selected five products from two categories—vacuum cleaners and mobile phones—on Trendyol, each with more than 2,000 reviews. These categories were chosen because they provide a large volume of customer reviews and are among the most frequently reviewed product groups on e-commerce platforms. Importantly, the aim was not to evaluate product performance but to test the models' generalizability on a dataset constructed from different products within the same categories used in Dataset 1. Reviews with 1 or 2 stars were categorized as negative, while those with 4 or 5 stars were categorized as positive. Reviews with 3 stars were excluded to avoid ambiguity in neutral sentiment. Irrelevant fields were removed during preprocessing. This process resulted in a dataset of 8,943 customer reviews, which was used in its entirety to evaluate and compare the classification performance of both models.

## 2.3 Fine-Tuning the Models

### 2.3.1 Fine-Tuning the First BERT Model

We used the pre-trained “dbmdz/bert-base-turkish-128k-uncased” model published by the MDZ Digital Library team. Using the model's tokenizer, we converted the comments from our first dataset into token IDs and obtained the corresponding attention masks.

The model was fine-tuned on Google Colab using GPU acceleration. Training was conducted with a batch size of 32 for 3 epochs. The AdamW optimizer with a learning rate of  $5e^{-5}$  was employed, and CrossEntropyLoss was used as the loss function. A TensorDataset consisting of token IDs, attention masks, and sentiment labels was created, and a DataLoader with random sampling was used to feed the data in each batch.

### 2.3.2 Fine-Tuning the Second BERT Model

We followed the same training steps for the second model and used the abovementioned parameters, using the dataset introduced in Section 2.2.2. Identical hyperparameters ensured that any performance differences were attributable to the datasets rather than training settings.

## 2.4 Chi-Square Tests

Four Chi-square tests for independence were conducted to assess the differences in sentiment classification results between the two models.

The first Chi-square test evaluated whether a statistically significant difference exists between Model 1 (based on ratings) and Model 2 (based on the semantic and syntactic features of review text) in classifying reviews as positive or negative, given that the data are categorical. This test was conducted on the full test dataset, which comprises 8,943 reviews (as described in Section 2.2.3).

Given that large sample sizes can lead to inflated significance (p-values quickly approaching zero), we also conducted a second Chi-square test using a random subset of 300 reviews. This addresses the issue known as the “p-value problem,” where statistical significance may not equate to practical significance in large datasets [8].

We created a ground truth dataset of 300 reviews to compare both models' classifications with human-annotated labels. The initial agreement rate between the two annotators (first and second authors) was 85%. Discrepancies were resolved through discussion to finalize the labels.

The third Chi-square test compared Model 1 predictions to the ground truth data. The fourth Chi-square test compared Model 2 predictions to the ground truth. These tests help assess how well each model aligns with human sentiment interpretation.

### 3. Results

#### 3.1 Experimental Results

The performance evaluation of the proposed BERT-based sentiment classification models was conducted using two distinct datasets: the first model was fine-tuned on the dataset described in Section 2.2.1, and the second was trained on manually labeled data introduced in Section 2.2.2. Both models were evaluated under positive and negative sentiment conditions using standard classification metrics, including precision, recall, F1-score, accuracy, macro-average, and weighted-average. The full comparison of these metrics is visualized in Figure 2.

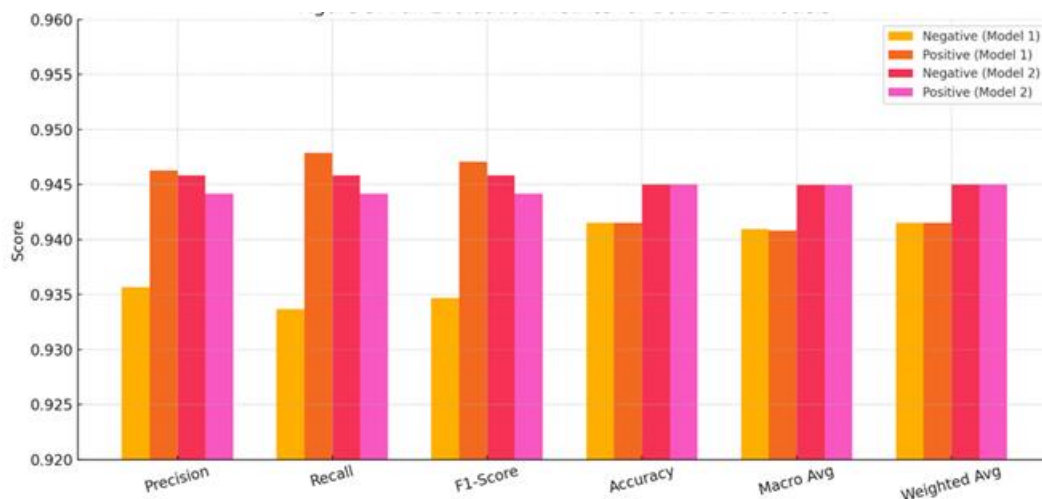


Figure 2. Full Evaluation Metrics for Both BERT Models across Sentiment Classes

As shown in Figure 2, the first model achieved an overall accuracy of 94.15%, with strong precision and recall values for both sentiment classes. However, it demonstrated a mild imbalance favoring the positive class. In contrast, with an overall accuracy of 94.5%, the second model exhibited near-identical precision, recall, and F1-score values across both classes. This indicates improved calibration and reduced bias, likely due to the use of manually labeled training data. The macro and weighted averages also reinforce this model's stability and robustness.

In the final phase, both fine-tuned BERT models were evaluated using the third dataset described in Section 2.2.3. Figure 3 comprehensively compares performance metrics—including precision, recall, F1-score, accuracy, macro average, and weighted average—across sentiment classes and models for this dataset.

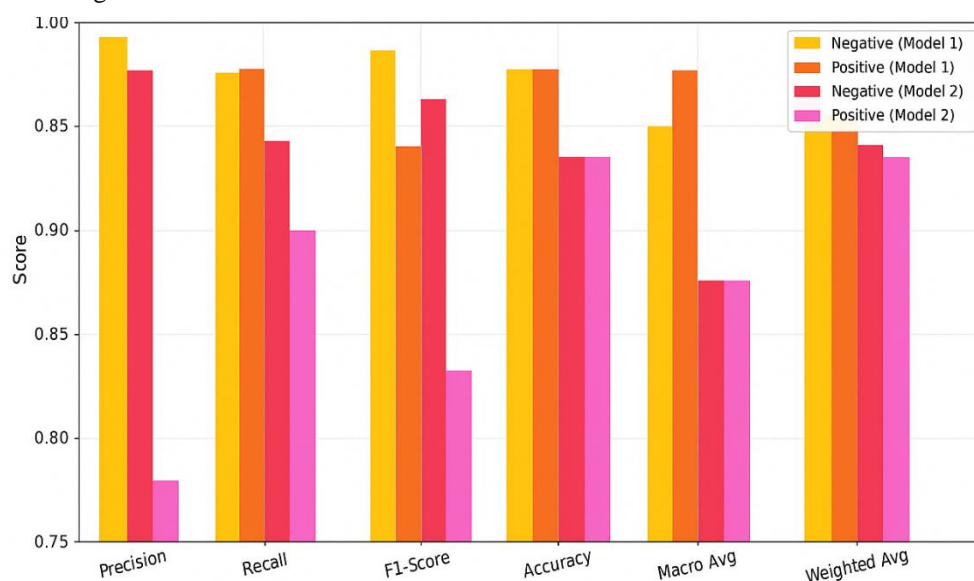


Figure 3. Evaluation of Both BERT Models on the Third Dataset.

The figure shows that the first model achieved an overall accuracy of 97.66%, with notably high performance on negative sentiment detection (precision: 0.995, recall: 0.976, F1-score: 0.986). While the precision for positive predictions is relatively

low (0.903), the recall for positive sentiment remains very strong (0.978), resulting in a solid F1-score of 0.939. These values indicate that the model generalizes well to unseen data, particularly in identifying negative sentiment with high confidence.

The second model, however, exhibits a notable decline in precision for the positive class (0.779) and a lower F1-score (0.836), although its recall for positive sentiment remains relatively high (0.901). The overall accuracy is 93.44%, and macro-level metrics such as macro average (0.878) and weighted average (0.936) reflect this drop in precision. These results indicate that the model struggles more with correctly identifying positive sentiment in the third dataset, likely due to the inherent mismatch between rating-based labels and actual textual sentiment expression.

Figure 4 highlights that while both models can handle sentiment classification tasks on the third data, the first model is better suited for generalization and precision-sensitive applications. Its superior balance across all metrics—particularly for negative sentiment—makes it a more robust option for deployment in real-world settings.

We also present the number of reviews classified as negative or positive by both BERT models using the dataset introduced in Section 2.2.3, which consists of 8,943 customer reviews. The classification results reveal how each model distributes predictions across sentiment categories, providing insight into their respective biases and generalization capabilities.

The first model classified 7,150 reviews as negative and 1,793 as positive. This means the model interpreted nearly 80% of the reviews as expressing negative sentiment, despite the original dataset having a slightly less skewed distribution. This disproportionate outcome suggests that Model 1, which was trained on star ratings, may be more sensitive to negative linguistic cues, or that the star-rating-based labeling approach overemphasized subtle dissatisfaction in text. As a result, this model tends to classify ambiguous or moderately critical comments as negative, possibly due to the nature of its training signals.

In contrast, the second model predicted 7,028 negative and 1,915 positive reviews. While still skewed toward negativity, this model showed a slightly more balanced classification than the first. The increased number of positive predictions implies that Model 2, trained on manually annotated text sentiment, may better grasp explicitly stated positivity—even when expressed in complex or nuanced forms. However, this broader recognition of positive sentiment may result in a slight decrease in precision, as observed in the quantitative evaluation metrics.

In summary, both models exhibit strong performance in sentiment classification, but their outputs reveal different tendencies. Model 1 is more conservative in identifying positive sentiment and may be biased toward negative classifications. Model 2, while more balanced in distribution, may struggle with precision when classifying positive reviews. These distinctions will be further examined in the discussion section in the context of labeling strategy, data imbalance, and the linguistic complexity of Turkish.

### 3.2 Chi-Square Results

In addition to standard performance metrics, we applied chi-square tests to further analyze whether the models' predictions and the human-labeled ground truth were statistically different.

The first test compared the full predictions of Model 1 and Model 2 over 8,943 samples and yielded a chi-square value of  $\chi^2 = 5.06$ ,  $p = 0.0240$ . Since  $p < 0.05$ , the result is statistically significant, indicating that the models differ in classification distributions. A second test, using a manually verified subset of 300 reviews as ground truth, showed an even stronger difference:  $\chi^2 = 9.8999$ ,  $p = 0.0016$ .

When the models were individually compared to the ground truth, Model 1 produced  $\chi^2 = 1.9003$ ,  $p = 0.1680$ , and Model 2 yielded  $\chi^2 = 3.3904$ ,  $p = 0.0656$ . Both values are statistically non-significant ( $p > 0.05$ ), indicating that neither model diverges substantially from the ground truth. However, Model 2's slightly higher value suggests a marginally greater deviation.

In summary, both models align reasonably well with the human-labeled data but differ significantly. This divergence suggests that, although both models are accurate, they exhibit distinct classification tendencies—potentially due to differences in the granularity of training data and labeling style.

## 4. Discussion

The central contribution of this study lies in examining the impact of two distinct labeling strategies: (i) automatic sentiment labeling derived from star ratings and (ii) manual sentiment labeling by human annotators. While star-based labels provide a scalable and easily accessible proxy for sentiment, they can at times misrepresent the underlying meaning of the text—for example, reviews rated with five stars that nevertheless contain negative expressions. By directly comparing these approaches, this study offers insights into the effectiveness of BERT-based models for classifying Turkish e-commerce product reviews. The results highlight the limitations of relying solely on star rating-based sentiment classification and demonstrate the advantages of using human-labeled textual data for more accurate sentiment detection.

### 4.1 Performance of BERT-Based Models in Turkish Review Classification

Model 1, trained on star ratings, achieved high recall and accuracy in detecting negative sentiment. However, it exhibited a consistent bias toward classifying reviews as negative. This pattern suggests that users may assign higher ratings even when they express dissatisfaction in the review text—resulting in a mismatch between numerical and textual sentiment [5], [21].

A striking example from our dataset illustrates this phenomenon clearly: as shown in Figure 5, a user gave a five-star rating

but used strongly negative and alarming language in the review text. This deliberate action was meant to ensure the complaint's visibility, as some platforms prioritize higher-rated reviews in display algorithms. Such discrepancies highlight the unreliability of star ratings as indicators of customer satisfaction.

Figure 4 summarizes these chi-square test results in a visual format.

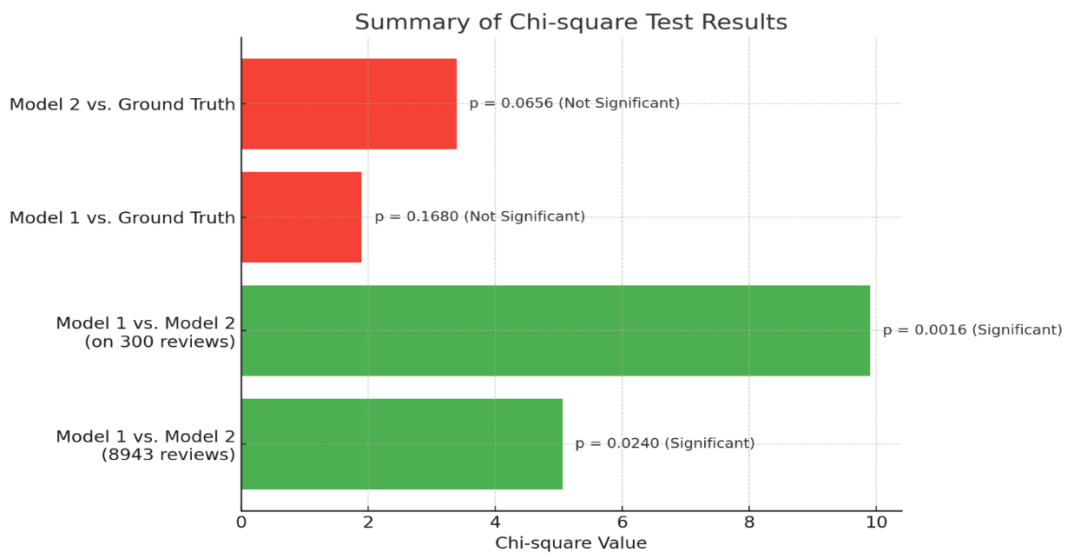


Figure 4. Summary of Chi-Square Test Results for Model Comparisons and Ground Truth Agreement



B\*\* Ö\*\* • 23 Kasim 2023 • Beden: M

Asla almayın yorum gözüksün diye 5 yıldız verdim. Her yanıma boyası çıktı, banyoya girmeyince çıkmadı. Kanser olmak istiyorsanız da siz bilirsiniz!!!

Never buy it. I gave 5 stars just so my comment would be visible. The dye came off on my whole body and didn't wash away until I showered. If you want to get cancer, that's your choice!!!

Figure 5. An illustrative review highlighting the misalignment between the numerical star rating and the sentiment expressed in the textual content.

Conversely, Model 2, trained on manually labeled textual sentiment, delivered a more balanced classification across sentiment categories. Despite slightly lower performance for positive sentiments, its predictions were more aligned with the nuanced intent behind customer feedback. Chi-square test results confirmed that while both models aligned similarly with the ground truth, their classification shapes decisions were significantly different—highlighting how the nature of training data (indirect stars vs. direct labels) shapes model behavior [26], [27].

#### 4.2 Linguistic Suitability of BERT for Turkish Sentiment Analysis

Turkish poses significant challenges to traditional sentiment analysis methods due to its agglutinative morphology, rich suffixation, and flexible word order. Dictionary- or rule-based systems fail to capture such linguistic complexity [12]. BERT enables deeper semantic comprehension through contextual embeddings, particularly domain-specific variants like dbmdz/bert-base-turkish-128k-uncased [19].

Several studies have confirmed the superiority of BERT over older models, such as CNN, LSTM, and GRU, in Turkish sentiment analysis [28]. BERT also outperforms multilingual models like XLM-RoBERTa and fastText, particularly in handling noisy, informal, and error-prone user-generated content [23], [29], [30].

### 4.3 Practical Implications for Businesses

Relying solely on star ratings can lead to skewed assessments of customer satisfaction. Text-based sentiment analysis offers more granular insights into consumer perceptions, enabling firms to identify hidden dissatisfaction patterns and respond effectively [12].

Integrating NLP into Customer Relationship Management (CRM) systems enhances responsiveness, informs product development, and strengthens brand reputation. Businesses in highly competitive sectors, such as fashion, electronics, and food delivery, may prioritize semantic sentiment analysis to retain customers and detect early warning signs of customer churn [31].

### 4.4 Societal and Cross-Sector Applications of Sentiment Analysis

The scope of AI-driven sentiment analysis extends far beyond e-commerce. In healthcare, it is used to analyze patient reviews, detect mental health symptoms through language use, and assess quality of care [28]. In finance, sentiment analysis helps gauge investor sentiment and predict stock trends.

Social media platforms utilize sentiment detection to combat misinformation, hate speech, and disinformation [32], [33], [34]. Governments can also analyze public sentiment in policy-making, fostering citizen-driven governance models [9].

### 4.5 Ethical Considerations and Responsible AI Development

Despite its advantages, AI-based sentiment analysis raises significant ethical concerns. Biased training data might perpetuate stereotypes or systematically disadvantage particular groups, especially in hiring or financial risk assessment [11]. Under-resourced languages, such as Turkish, face additional risks due to the limited and unbalanced datasets [35].

Responsible AI requires anonymized data, transparency reports, human oversight, and compliance with privacy regulations, such as KVKK [27], [31], [36]. Model interpretability and ethical governance must be central to any large-scale NLP deployment.

In the present study, the training datasets were constructed in a randomized and balanced manner, which minimizes the risk of systematic bias. Since the primary aim was to compare star rating-based labeling with human-labeled sentiment data, rather than to model specific user demographics or product categories, the dataset does not carry the social or demographic biases often observed in other applications. Nevertheless, future research should focus on creating larger and more diverse human-labeled datasets for Turkish sentiment analysis. Such datasets would not only strengthen model robustness but also provide a stronger safeguard against potential biases that may emerge when scaling sentiment analysis applications to broader domains.

### 4.6 The Future of Work: Human-AI Collaboration in Sentiment Analysis

AI is transforming work, particularly in text analytics and customer experience management. Automation reduces the burden of repetitive tasks, but it must not eliminate the role of human judgment. Hybrid architectures, such as human-in-the-loop (HITL) systems, ensure accountability and maintain quality by validating AI predictions with expert input [7], [9], [36].

Future research should explore how emotional intelligence, cultural sensitivity, and contextual reasoning—strengths still unique to human cognition—can be embedded within AI workflows. This will help sustain ethical and informed decision-making in an increasingly automated world.

Looking ahead, future research should prioritize three key areas. First, developing larger and more balanced Turkish sentiment datasets is essential for improving model generalizability and fairness. Second, exploring multimodal sentiment analysis that integrates textual, visual, and auditory inputs can lead to a more holistic understanding of user emotions and opinions. Third, there is a growing need to design human-AI collaboration frameworks that support transparent, explainable, and ethically aligned decision-making.

By harnessing the capabilities of advanced natural language processing, organizations can gain deeper insights into customer experiences, optimize decision-making processes, and uphold principles of fairness, accountability, and trust when deploying AI systems.

## 5. Conclusion

This study demonstrates that fine-tuned BERT models significantly improve sentiment classification accuracy in Turkish e-commerce reviews compared to traditional star-rating approaches. While Model 1 demonstrated high performance in detecting negative sentiment, its reliance on numerical ratings resulted in misclassifications due to misaligned sentiment. In contrast, Model 2, trained on directly labeled textual sentiment, provided a more balanced yet slightly less precise classification outcome.

Our findings reveal several important insights regarding developing and applying sentiment analysis models in Turkish. First and foremost, the nature of training data plays a crucial role in shaping model behavior. The method used to label sentiment, whether based on star ratings or direct textual annotation, can lead to significantly different classification patterns. This highlights the importance of carefully curating and balancing datasets, especially when dealing with subjective content such as customer feedback.

Second, BERT has proven to be particularly well-suited for the Turkish language, which poses unique linguistic challenges due to its agglutinative morphology and flexible syntax. This confirms the advantage of using language-adapted architectures in under-resourced or morphologically rich languages.

Third, the practical importance of text-based sentiment analysis extends far beyond e-commerce. In business, sentiment tools can enhance customer relationship management and market strategy formulation. In broader societal contexts, AI-powered sentiment analysis is applied in healthcare to assess patient satisfaction and mental health indicators, and in public policy to understand citizen feedback and opinions. These applications demonstrate the transformative potential of sentiment technologies when implemented responsibly.

## References

- [1] M. S. Akin, "Enhancing e-commerce competitiveness: A comprehensive analysis of customer experiences and strategies in the Turkish market," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 1, p. 100222, Mar. 2024, doi: 10.1016/j.joitmc.2024.100222.
- [2] N. Yücel and Ö. Cömert, "Müşteri Duyarlılığını Keşfetmek İçin Yapay Zeka Destekli Analiz ile Çevrimiçi Ürün İncelemelerinden Anlamlı Bilgiler Elde Etme," *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 35, no. 2, pp. 679–690, Sep. 2023, doi: 10.35234/fumbd.1305932.
- [3] M. T. Barutcu and B. Basak, "Customer Complaints about E-Commerce Sites: Content Analysis," 2018.
- [4] Elif Ayanoğlu, Zeynep Çolak, Toygar Tanyel, Hasan Yunus Sarioğlu, and Banu Diri, "Detection and Classification of Customer Comments Containing Complaints," Nov. 2023, doi: 10.5281/ZENODO.10254498.
- [5] M. Demircan, A. Seller, F. Abut, and M. F. Akay, "Developing Turkish sentiment analysis models using machine learning and e-commerce data," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 202–207, Jun. 2021, doi: 10.1016/j.ijcce.2021.11.003.
- [6] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 6, p. 100059, Mar. 2024, doi: 10.1016/j.nlp.2024.100059.
- [7] A. Ocal, "Perceptions of the Future of Artificial Intelligence on Social Media: A Topic Modeling and Sentiment Analysis Approach," *IEEE Access*, vol. 12, pp. 182386–182409, 2024, doi: 10.1109/ACCESS.2024.3510526.
- [8] A. Ocal, "Framing, Emotions, Salience: The Future of AI as Seen by Redditors," Ph.D., Syracuse University, United States -- New York, 2023. Accessed: Oct. 19, 2023. [Online]. Available: <https://www.proquest.com/docview/2845416849>
- [9] A. Ocal and K. Crowston, "Framing and feelings on social media: the futures of work and intelligent machines," *ITP*, vol. 37, no. 7, pp. 2462–2488, Apr. 2024, doi: 10.1108/ITP-01-2023-0049.
- [10] A. Öcal, L. Xiao, and J. Park, "Reasoning in social media: insights from Reddit 'Change My View' submissions," *Online Information Review*, vol. 45, no. 7, pp. 1208–1226, Jan. 2021, doi: 10.1108/OIR-08-2020-0330.
- [11] A. Ocal, "Perceptions of AI Ethics on Social Media," in *2023 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*, IEEE, 2023. doi: 10.1109/ETHICS57328.2023.10155069.
- [12] S. F. Yilmaz, E. B. Kaynak, A. Koc, H. Dibeklioglu, and S. S. Kozat, "Multi-Label Sentiment Analysis on 100 Languages With Dynamic Weighting for Label Imbalance," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 34, no. 1, pp. 331–343, Jan. 2023, doi: 10.1109/TNNLS.2021.3094304.
- [13] İ. Yurtseven, S. Bagriyanik, and S. Ayvaz, "A Review of Spam Detection in Social Media," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, Sep. 2021, pp. 383–388. doi: 10.1109/UBMK52708.2021.9558993.
- [14] S. Ayvaz, S. Yıldırım, and Y. B. Salman, "Türkçe Duygu Kütüphanesi Geliştirme: Sosyal Medya Verileriyle Duygu Analizi Çalışması," *European Journal of Science and Technology*, pp. 51–60, Aug. 2019, doi: 10.31590/ejosat.537085.
- [15] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics*, vol. 35, no. 1, pp. 136–147, Apr. 2018, doi: 10.1016/j.tele.2017.10.006.
- [16] S. M. Mohammad and P. D. Turney, "Crowdsourcing A Word–Emotion Association Lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, Aug. 2013, doi: 10.1111/j.1467-8640.2012.00460.x.
- [17] W. van Atteveldt, M. A. C. G. van der Velden, and M. Boukes, "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms," *Communication Methods and Measures*, vol. 15, no. 2, pp. 121–140, Apr. 2021, doi: 10.1080/19312458.2020.1869198.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [19] U. U. Acikalın, B. Bardak, and M. Kutlu, "Turkish Sentiment Analysis Using BERT," in *2020 28th Signal Processing and Communications Applications Conference (SIU)*, Gaziantep, Turkey: IEEE, Oct. 2020. doi: 10.1109/siu49456.2020.9302492.

- [20] B. Ciftci and M. S. Apaydin, "A Deep Learning Approach to Sentiment Analysis in Turkish," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey: IEEE, Sep. 2018, pp. 1–5. doi: 10.1109/idap.2018.8620751.
- [21] G. Yavuz, "Web Kazıma Ve Duygu Analizi Temelli Ürün Analiz Sistemi," 2023. *Master's Thesis*.
- [22] M. Masarifoglu *et al.*, "Sentiment Analysis of Customer Comments in Banking using BERT-based Approaches," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, Istanbul, Turkey: IEEE, Jun. 2021, pp. 1–4. doi: 10.1109/siu53274.2021.9477890.
- [23] M. Arzu and M. Aydoğan, "Türkçe Duygu Sınıflandırma İçin Transformers Tabanlı Mimarilerin Karşılaştırılması Analizi," *JCS*, Aug. 2023, doi: 10.53070/bbd.1350405.
- [24] N. Paker and B. Kizilirmak, "Çevrim İçi Müşteri Yorumlarını Etkileyen Faktörler Üzerine Keşifsel Bir Çalışma: Trendyol Örneği," *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, vol. 23, no. 4, pp. 1393–1414, Dec. 2023, doi: 10.18037/ausbd.1309934.
- [25] S. İlhan Omurca, E. EkiNci, E. Yakupoğlu, E. Arslan, and B. Çapar, "Automatic Detection of the Topics in Customer Complaints with Artificial Intelligence," *Balkan Journal of Electrical and Computer Engineering*, vol. 9, no. 3, pp. 268–277, Jul. 2021, doi: 10.17694/bajece.832274.
- [26] B. Teke, S. N. Yazıcı, G. Zamir, A. B. Budak, and I. Karabey Aksakallı, "BERTurk-Based Sentiment Analysis on E-Commerce Multi Domain Product Reviews," *Afyon Kocatepe University Journal of Sciences and Engineering*, vol. 25, no. 3, pp. 497–509, May 2025, doi: 10.35414/akufemubid.1537513.
- [27] S. Yildirim, "Fine-tuning Transformer-based Encoder for Turkish Language Understanding Tasks," Jan. 30, 2024, *arXiv*: arXiv:2401.17396. doi: 10.48550/arXiv.2401.17396.
- [28] P. Savci and B. Das, "Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 3, pp. 227–237, Mar. 2023, doi: 10.1016/j.jksuci.2023.02.017.
- [29] S. Sinha, A. Jayan, and R. Kumar, "An Analysis and Comparison of Deep-Learning Techniques and Hybrid Model for Sentiment Analysis for Movie Review," in *2022 3rd International Conference for Emerging Technology (INCET)*, May 2022, pp. 1–5. doi: 10.1109/INCET54531.2022.9824630.
- [30] T. Tanyel, B. Alkurdi, and S. Ayvaz, *Linguistic-based Data Augmentation Approach for Offensive Language Detection*. 2022, p. 6. doi: 10.1109/UBMK55850.2022.9919562.
- [31] M. Gürbüz and M. Kotan, "Multi-Category E-Commerce Insights via Social Media Analysis using Machine Learning and BERT," *acin*, vol. 0, no. 0, pp. 0–0, Feb. 2025, doi: 10.26650/acin.1483488.
- [32] A. Dalgali and K. Crowston, "Sharing Open Deep Learning Models," *presented at the Hawai'i International Conference on System Science.*, Jan. 2019. doi: 10.24251/HICSS.2019.256.
- [33] A. Dalgali and K. Crowston, "Algorithmic Journalism and Its Impacts on Work," *Computation + Journalism Symposium*, 2020, [Online]. Available: <https://cj2020.northeastern.edu/>
- [34] A. Dalgali and K. Crowston, "Factors Influencing Approval of Wikipedia Bots," in *The Hawaii International Conference on System Sciences*, 2020, p. 10. [Online]. Available: <http://hdl.handle.net/10125/63757>
- [35] M. Bozuyula, "Sentiment Analysis of Turkish Drug Reviews with Bidirectional Encoder Representations from Transformers," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–17, Jan. 2024, doi: 10.1145/3626523.
- [36] H. A. Love *et al.*, "The Future of Work in the Age of Automation: Proceedings of a Workshop on Norbert Wiener's 21st Century Legacy," *IEEE Trans. Technol. Soc.*, pp. 1–23, 2024, doi: 10.1109/TTS.2024.3476041.

## Article Information Form

### Acknowledgments

Thanks to Emre Dalcı for his support during the data collection and analysis phases.

### Conflict of Interest Notice

The author declares that there is no conflict of interest regarding the publication of this paper.

### Ethical Approval

It is declared that during the preparation process of this study, scientific and ethical principles were adhered to, and all studies cited are listed in the bibliography.

### Availability of data and material

Not applicable

**Artificial Intelligence Statement**

The author utilized ChatGPT to enhance the clarity and grammatical accuracy of the manuscript. This tool was used as a language aid during the writing and revision process.

**Plagiarism Statement**

This article has been scanned by Turnitin.