Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1836772

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 2: 826-834 / March 2026

GREEN SENTIMENT ANALYSIS IN E-COMMERCE REVIEWS: A COMPARATIVE MACHINE LEARNING APPROACH ON TURKISH CONSUMER FEEDBACK

Cevher ÖZDEN[1]*

[1]*Cukurova University, Faculty of Arts and Sciences, Department of Computer Sciences, 01250, Adana, Türkiye*

Abstract: This study investigates the effectiveness of three sentiment classification approaches, i.e. Logistic Regression (LR), Support Vector Classification (SVC), and a fine-tuned BERTurk model, on Turkish e-commerce reviews related to environmentally conscious, or "green," products. Using a real-world dataset drawn from Trendyol, one of Türkiye's largest online marketplaces, we preprocessed and filtered the data to focus on user-generated product comments that reference sustainability-oriented themes. Each model was evaluated using standard classification metrics, including accuracy and macro-averaged F1-score, to assess both overall performance and sensitivity to class imbalance. The results show that while classical machine learning methods such as LR and SVC provide reasonably high accuracy, they struggle to distinguish neutral sentiment effectively, which is an issue commonly encountered in Turkish-language sentiment tasks. In contrast, the BERTurk model achieved the highest overall performance, with an accuracy of 0.91 and a macro F1-score of 0.67. It was particularly effective in detecting positive and negative sentiment, while still exhibiting the known difficulty of identifying neutral expressions. These findings suggest that transformer-based models offer a clear advantage in extracting sentiment from morphologically rich languages like Turkish, especially in domains where emotional nuance and linguistic ambiguity are prevalent. The study contributes to both the sentiment analysis literature and Management Information Systems research by demonstrating the value of domain-specific deep learning for consumer analytics in green commerce. It highlights practical implications for businesses aiming to understand and respond to public attitudes toward sustainable products and emphasizes the need for improved modeling of neutral sentiment. Future work should focus on expanding Turkish sentiment datasets, addressing class imbalance, and refining model architectures to better capture the subtleties of eco-conscious consumer expression.

Keywords: Turkish sentiment analysis, BERTurk, Green commerce, E-commerce reviews, Transformer models

***Corresponding author:** Cukurova University, Faculty of Arts and Sciences, Department of Computer Sciences, 01250, Adana, Türkiye
E mail: cozden@cu.edu.tr (C. ÖZDEN)
Cevher ÖZDEN https://orcid.org/0000-0002-8445-4629

Received: December 05, 2025
Accepted: February 19, 2026
Published: March 15, 2026

Cite as: Özden, C. (2026). Green sentiment analysis in e-commerce reviews: a comparative machine learning approach on Turkish consumer feedback. *Black Sea Journal of Engineering and Science*, *9*(2), 826-834.

1. Introduction

E-commerce has turned customer reviews into one of the most important data sources for both researchers and practitioners. Platforms such as Trendyol, Amazon and Alibaba host millions of user-generated comments that record everyday experiences with products and services. For management information systems (MIS), these texts are not only "opinions" but also a continuous input to decision-support systems, business intelligence dashboards and recommendation engines. Online reviews influence purchase decisions, shape brand reputation and guide operational changes, especially when they are systematically analyzed using sentiment analysis and other natural language processing (NLP) techniques (Huang et al., 2023; Daza et al., 2024).

At the same time, sustainability has become a strategic concern in digital commerce. Research on sustainable or "green" e-commerce shows that environmental practices such as eco-friendly packaging, carbon-neutral shipping or sustainable sourcing can support competitiveness and customer loyalty (Oláh et al., 2022). Policy reports and academic reviews highlight that B2C e-commerce has a complex environmental footprint, with both positive and negative effects depending on logistics, returns and consumer behaviour (Mangiaracina et al., 2015; UNCTAD, 2024). In this debate, the voice of the online consumer is still under-used: most studies focus on emissions and logistics models rather than on what customers actually say about "green" attributes in their everyday shopping.

Within MIS, Green Information Systems (Green IS) research argues that information systems can support pro-environmental behaviour and more sustainable decision-making at individual, organisational and societal levels (Brendel et al., 2022; Singh et al., 2022). However, much of the Green IS literature is conceptual or organisation-centric. It looks at energy-efficient infrastructures, eco-dashboards or conference systems, rather than large-scale analyses of consumer-generated reviews in commercial platforms. As a result, we know comparatively little about how "green sentiment" is

expressed in real-world e-commerce reviews, especially in non-English markets.

In parallel, sentiment analysis on online reviews has grown into a mature research area at the intersection of MIS, computer science and marketing. Recent reviews show a rapid expansion of machine-learning and deep-learning approaches for opinion mining in e-commerce, often with explicit links to decision support and managerial use (Daza et al., 2024; Huang et al., 2023; Doğan and Kara, 2025; Aguilar-Moreno et al., 2024). Sentiment outputs are used to rank products, support product selection, evaluate service quality, and feed decision-support dashboards for managers (Nave et al., 2018; Wu et al., 2024; Zhang et al., 2022). Social media analytics and online review mining are now recognised MIS tools that help organisations understand consumer attitudes and adjust strategies more quickly (Loke and Pathak, 2023; Macías Urrego et al., 2024).

More recently, a smaller but growing body of work has started to focus specifically on eco-friendly products and sustainability-related reviews. Maarif et al. (2024) analysed Amazon reviews for eco-friendly products to extract sustainability-related themes using sentiment analysis and topic modelling. Shaik Vadla et al. (2024) applied BERT and T5 models to identify sentiment towards eco-friendly products and discussed how these insights could support greener product design. Similar work shows that AI-based sentiment analysis can provide additional sustainability insights that complement life-cycle assessment data (Maarif et al., 2024). These studies clearly demonstrate the potential of "green sentiment analytics". Yet they are mostly based on global platforms such as Amazon and are dominated by English-language or international data.

The Turkish e-commerce ecosystem is large and fast-growing, reaching a trade volume of 3.16 trillion TRY, with Trendyol as one of its key actors serving over 30 million active customers (Bilik, 2023). However, academic work on Turkish-language online reviews is still limited. Existing studies often concentrate on generic sentiment classification in Turkish, sometimes using BERTurk or other transformer-based models, but they rarely isolate environmentally framed products or green attributes in their analyses. Recent Turkish research on sentiment analysis in e-commerce platforms confirms the importance of this topic but still treats sustainability only indirectly or not at all (Savci and Das, 2023; Doğan and Kara, 2025). To our knowledge, there is no large-scale MIS-oriented study that combines (i) a real-world Turkish retail platform, (ii) a "green" subset of products and reviews, and (iii) a direct comparison of classical machine-learning models with a domain-adapted transformer model.

This paper addresses that gap by focusing on green sentiment analysis in Trendyol product reviews from a management information systems perspective. Using the publicly available Kaggle dataset "Trendyol Product Comments" (Demir, 2024), which contains tens of thousands of product reviews and metadata, we build a domain-specific subset of "green" comments by filtering for environmentally related keywords such as *organik*, *sürdürülebilir*, *doğal* and *pamuk* in both product names and review texts. We then apply a multi-step pipeline: text cleaning and normalisation, rule-based sentiment labelling from star ratings, classical sentiment modelling with TF-IDF + Logistic Regression / Support Vector Machines, and a fine-tuned transformer model based on a Turkish BERT architecture. This allows us to compare traditional and deep-learning approaches on the same MIS-relevant task.

From an MIS viewpoint, our goal is twofold. First, we want to describe how Turkish consumers evaluate products that carry explicit or implicit "green" cues in their descriptions, and whether these evaluations are predominantly positive, negative or neutral. Understanding this distribution helps both researchers and practitioners assess whether green claims are perceived as credible and valuable, or as mere marketing noise. Second, we aim to demonstrate how an integrated sentiment-analysis pipeline can be used as a decision-support component in a broader Green IS context: managers could monitor "green sentiment" over time, compare brands or categories, and identify recurring themes in positive or negative eco-related feedback, feeding into product design, sourcing and communication decisions (Ramos et al., 2023; Brendel et al., 2022; Chiang, 2024).

This study makes three main contributions. First, it provides one of the first large-scale analyses of green-related consumer sentiment in Turkish e-commerce, using real Trendyol data. Second, it compares classical machine-learning models and a BERT-based transformer model in this specific context, offering practical guidance for MIS practitioners who wish to operationalise green sentiment analytics. Third, it connects sentiment analysis with Green IS and sustainable e-commerce literature, showing how user-generated "green" reviews can be embedded into decision-support systems for more environmentally aware digital retail strategies.

2. Materials and Methods

2.1. Materials

The empirical analysis is based on the publicly available Trendyol Product Comments dataset, published on Kaggle as a SQLite database (TrendyolProduct.sqlite3). The database contains user reviews collected from Trendyol, one of the largest business-to-consumer e-commerce platforms in Türkiye. It consists of two core tables:

- TBL_Product, which stores product-level metadata (product ID, product name, brand, and product URL).
- TBL_Comment, which stores review-level data (comment ID, product ID, free-text review, and a 1–5 star rating).

Using an inner join on the Product_Id field, we combined both tables into a single dataset where each row represents one review together with its associated product information. In total, this initial merge yielded 43,923 reviews.

Construction of the "green" review subset; The goal of this study is to focus on reviews that explicitly or implicitly refer to environmentally related product features. To operationalise this, we constructed a lexicon of "green" keywords in Turkish, based on common terms used in sustainability and textile marketing. The following stems and expressions were used:

- organik (organic)
- sürdürülebilir (sustainable)
- geri dönüştür (recycled / recyclable)
- eko (eco-, ecological, eco-friendly)
- doğa dostu (nature-friendly)
- doğal (natural)
- pamuk (cotton; especially %100 pamuk / cotton)

These terms were searched (case-insensitive, using regular expressions) in both the product name (Product_Name) and the review text (Comment_Content). A review was included in the "green" subset if *either* the product name *or* the review text contained at least one of the above patterns.

After applying this filter, we obtained 5,293 reviews, which constitute the core dataset used in the rest of the analysis. This subset captures products that are marketed with green cues (e.g., "%100 pamuk", "organik pamuklu") as well as reviews where users themselves refer to natural or eco-related aspects.

Data pre-processing; All text processing and modelling were carried out in Python. The review text used for modelling is the Comment_Content field, and we also keep Product_Name for descriptive analyses.

We implemented a simple but robust pre-processing pipeline tailored to Turkish text:

1. Encoding corrections: Because the raw SQLite export occasionally contained mis-encoded Turkish characters, we manually corrected the most frequent patterns using string replacement.
2. Lowercasing: All text was converted to lower case to reduce sparsity (e.g., "Güzel" and "güzel" are treated as the same token).
3. Removal of non-alphabetic characters: We retained only Turkish letters (a–z, ç, ğ, ı, ö, ş, ü) and whitespace, removing digits, punctuation and other symbols using a regular expression. This step keeps the focus on lexical content.
4. Whitespace normalization: Multiple spaces collapsed into a single space and leading/trailing spaces were trimmed.

The output of this step is stored in a new field called Cleaned_Content, which serves as the main input for all downstream sentiment models. For exploratory keyword and frequency analysis, we additionally tokenised Cleaned_Content into word lists and removed Turkish stop words using the NLTK Turkish stopword list; however, the classification models operate directly on the cleaned text.

Rating-based sentiment labelling; Each review in the original database includes a 1–5 star rating (Comment_Evaluation). We converted this ordinal scale into a three-class sentiment label as follows:

- Positive: ratings of 4 or 5
- Neutral: rating of 3
- Negative: ratings of 1 or 2

This rule-based mapping is commonly used in review mining and yields a straightforward ground truth for supervised learning. After mapping, the class distribution in the green review subset was:

- Positive: 4,479 reviews (84.62%)
- Negative: 443 reviews (8.37%)
- Neutral: 371 reviews (7.01%)

This distribution reveals a strong skew towards positive sentiment, which we explicitly account for when choosing modelling techniques and evaluation metrics.

The resulting labelled dataset, containing Cleaned_Content and the categorical sentiment label, was exported as trendyol_yesil_yorumlar_labeled.csv for reuse in all experiments.

Train–test split and evaluation metrics; To evaluate model performance on unseen data, we split the labelled dataset into training and test sets: 80% of the reviews were used for training and 20% were held out for testing. The split was stratified by sentiment label to preserve the original class proportions in both subsets. The random seed was fixed to ensure reproducibility. Given the pronounced class imbalance, we report not only overall accuracy, but also on macro-averaged F1 (simple average of F1 across classes), weighted F1 (F1 weighted by class support), and class-specific precision, recall, and F1-scores, especially for the minority *negative* and *neutral* classes. These metrics allow a more nuanced assessment than accuracy alone.

2.2. Methods

As a first step, we established classical machine-learning baselines using TF–IDF features combined with linear classifiers implemented in scikit-learn.

Text representation (TF–IDF); We vectorised Cleaned_Content using a TF–IDF representation with the following configuration:

- n-gram range: unigrams and bigrams ((1, 2)), capturing single words and short phrases (e.g., "çok güzel");
- Minimum document frequency: min_df = 5, so only terms appearing in at least five reviews were kept;
- Default scikit-learn tokenisation, operating on whitespace-separated tokens.

The TF–IDF vectoriser was fitted on the training set only and then applied to both train and test sets to avoid information leakage.

Logistic Regression; We trained a multinomial Logistic Regression classifier with max_iter = 1000 to ensure

convergence on the high-dimensional TF–IDF space; class_weight = "balanced" to compensate for the under-representation of negative and neutral reviews; default regularisation (L2) and one-vs-rest scheme for multiclass classification. This model serves as a strong linear baseline and is widely used in text classification.

Support Vector Classifier (SVC); As a second baseline, we trained a Support Vector Classifier with a linear kernel, balanced class_weight, and other hyperparameters kept at scikit-learn defaults. SVC is known to perform well on sparse, high-dimensional feature spaces such as TF–IDF. Both classical models were evaluated on the held-out test set using the metrics described in Section 3.5.

Transformer-based deep learning model (BERTurk); To capture richer linguistic context and evaluate the added value of deep learning, we fine-tuned a Turkish BERT model on the same task. We used the dbmdz/bert-base-turkish-cased checkpoint from the HuggingFace model hub, which is a cased BERT model pretrained on large-scale Turkish text.

Tokenisation and dataset preparation; The BERTurk tokenizer was applied to the Cleaned_Content field with 128 maximum tokens sequence length, max_length padding and truncation set to true for longer reviews. The labelled pandas DataFrame was converted to HuggingFace Dataset objects for the training and test splits. Tokenised datasets were then stored in PyTorch tensor format, including input_ids, attention_mask, and the numeric label.

Model configuration and training; We instantiated AutoModelForSequenceClassification with num_labels = 3 corresponding to *negative*, *neutral* and *positive* and id2label / label2id dictionaries for readability of outputs. The classification head on top of BERT was randomly initialised (as indicated by the standard warning from the library) and trained jointly with the base model. Training was performed using the HuggingFace Trainer API with the following TrainingArguments:

- Epochs: num_train_epochs = 3;
- Learning rate: 2e-5;
- Batch size: per_device_train_batch_size = 16, per_device_eval_batch_size = 32;
- Weight decay: 0.01;
- Logging: every 50 steps;
- Evaluation performed on the held-out test set at the end of training.

No additional pretraining or domain adaptation was performed beyond fine-tuning on the green Trendyol subset.

Evaluation; After training, we evaluated the BERTurk model on the same test split used for the classical baselines. Predictions were obtained via Trainer.predict, and class probabilities were converted to labels via argmax over the logits. We then computed overall accuracy, macro- and weighted F1-scores, class-wise precision, recall, F1 as well as confusion matrix to inspect typical misclassifications, especially between neutral and neighbouring classes.

These results are reported and compared to the TF–IDF + Logistic Regression / SVC baselines in the Results section. The analyses were conducted on the final dataset of 5,293 “green” Trendyol reviews. This section first presents the descriptive sentiment distribution, followed by the results of the baseline models and the fine-tuned BERTurk model. All experiments used the same stratified train–test split.

3. Results

The sentiment distribution of the filtered “green” subset is presented in Table 1. As shown, the dataset is dominated by positive reviews, which account for more than four-fifths of all observations. Neutral and negative reviews together comprise a small minority. This imbalance reflects the general trend in user-generated product feedback, where satisfied customers are more likely to leave comments, especially for textile products marketed with eco-friendly language.

Table 1. Sentiment distribution in the green review subset

Sentiment Class	Count	Percentage
Positive	4,479	84.62%
Neutral	371	7.01%
Negative	443	8.37%
Total	5,293	100%

This uneven distribution has two methodological implications. First, classification models naturally learn the positive class more easily, since its linguistic markers (e.g., *güzel*, *rahat*, *kaliteli*) appear frequently. Second, minority classes, especially the neutral category, are more difficult to model due to their limited examples and less distinctive lexical patterns. The first benchmark consisted of Logistic Regression and Support Vector Classification trained on TF-IDF representations. Their comparative performance is summarised in Table 2. Logistic Regression achieved an accuracy of 0.82, while SVC reached 0.83. In both models, the positive class was predicted with high precision and recall, whereas neutral and negative classes showed considerably weaker results. This behaviour was expected given the sparse nature of TF–IDF vectors and the skewed class distribution.

Although both classical models performed reliably on the majority class, they struggled to capture the more nuanced expressions associated with neutral sentiment. Many neutral reviews were absorbed into the positive class. Fine-tuning the BERTurk model produced the strongest results among all tested approaches. The model reached an accuracy of 0.91 and achieved substantial gains in macro-averaged F1, indicating better treatment of minority classes. Class-wise metrics for BERTurk are presented in Table 3. The quantitative assessment of

model performance, illustrated in Figure 1, demonstrates the significant superiority of the transformer-based BERTurk architecture over traditional machine learning baselines (Logistic Regression and SVC). While all models achieved competitive accuracy scores largely due to the dominance of the positive class, the F1-macro metric reveals a critical distinction in model robustness. BERTurk achieved a markedly higher F1-macro score compared to LogReg and SVC, indicating its enhanced capacity to generalize across minority classes in an imbalanced dataset.

Table 2. Overall model performance on test set

Model	Accuracy	F1-macro	F1-weighted
Logistic Regression (TF–IDF)	0.82	0.56	0.80
Support Vector Classifier (TF–IDF)	0.83	0.58	0.81
BERTurk (fine-tuned)	0.91	0.67	0.90

Table 3. BERTurk – Class-wise Precision, Recall, and F1

Class	Precision	Recall	F1-score	Support
Negative	0.69	0.81	0.75	89
Neutral	0.38	0.24	0.30	74
Positive	0.96	0.97	0.96	894
Macro avg	0.68	0.67	0.67	—
Weighted avg	0.90	0.91	0.90	—

The BERT-based model was especially effective at detecting negative reviews, improving their F1-score by a large margin compared to the TF–IDF models. This suggests that contextual embeddings capture subtle dissatisfaction markers in Turkish better than sparse lexical features. Nevertheless, the neutral class remained challenging; even with BERT, many neutral statements were labelled as positive. This may stem from rating–text mismatches or from the inherently ambiguous nature of neutral expressions. The confusion matrix further illustrates common misclassification patterns and is provided in Figure 2. The matrix shows that neutral reviews are primarily misclassified as positive, whereas negative reviews are more cleanly separated. This supports the textual analysis: negative comments tend to include clearer emotional markers, while neutral ones often consist of short, content-light statements. Overall, the results indicate that transformer-based modelling provides a clear improvement over classical approaches for sentiment analysis in Turkish e-commerce reviews. Even under severe class imbalance, BERTurk maintained superior accuracy and considerably stronger recall for negative sentiment. For MIS and e-commerce applications, this improvement is meaningful: capturing negative sentiment more accurately can uncover issues in eco-labelled products that would have been overlooked by simpler models. The confusion matrix shows that BERTurk achieves highly accurate predictions for positive and negative reviews, while neutral comments remain the most challenging category. Most neutral reviews are misclassified as positive, reflecting both the scarcity of neutral training examples and the inherently ambiguous linguistic structure of neutral expressions.

4. Discussion

The superior performance of the fine-tuned BERTurk model in our experiments is consistent with recent findings in Turkish sentiment analysis, especially in e-commerce contexts. For example, Özmen and Gündüz (2025) report that on Trendyol cosmetic product reviews, a lexicon-based SVM achieved over 93% accuracy. Similarly, Teke et al. (2025) found that among traditional classifiers, SVM reached about 88% accuracy on multi-category Trendyol reviews, whereas a BERTurk model achieved 95%. In our study the SVM (or Support Vector Classifier) and logistic regression models performed in the mid-80% range (accuracy ≈0.88), comparable to these results. In contrast, the transformer-based BERTurk achieved 0.91 accuracy, underscoring the advantage of contextual deep models. This advantage aligns with the literature: Öcal (2025) emphasizes that transformer models like BERT better capture Turkish's complex morphology and syntax than traditional approaches, which likely explains our model's improved capture of sentiment nuances. In sum, consistent with prior work, we find that TF–IDF-based LR/SVM methods yield strong but somewhat lower accuracy (often high-80s%) whereas pretrained BERTurk variants yield the highest accuracy (around 90%+) on Turkish product reviews.

A notable challenge in our results is the lower performance on neutral reviews. This mirrors a common finding in Turkish sentiment research that neutral sentiment is hard to classify. Zümberoğlu et al. (2025) note that "the definitive classification of neutral sentiment poses greater challenges," often leading to neutral examples being misclassified. They further observe that many Turkish sentiment datasets are imbalanced with far fewer neutral examples, which "exacerbates the difficulty of distinguishing the neutral class". Likewise, Teke et al. (2025) report that in their Trendyol dataset the neutral class had substantially lower precision, recall, and F1-score than the positive and negative classes. In our case, the macro-averaged F1 of 0.67 (versus 0.91 accuracy) reflects that the neutral class was indeed a minority and more ambiguous category, dragging down the average. These observations highlight that class imbalance and sentiment ambiguity (especially for neutral or mixed opinions) remain major challenges. Several studies advocate balancing the

dataset or augmenting neutral examples to mitigate this issue.

The findings of the study align with recent studies emphasizing the importance of category-based specialization in the analysis of e-commerce reviews. For instance, Gürbüz and Kotan (2025) demonstrated using Trendyol data that employing category-based analytical frameworks (multi-category insights), rather than generic sentiment analysis, proves more effective in understanding sectoral dynamics. This supports the methodological validity of the 'green product' focused filtering approach applied in our study. Furthermore, the data augmentation approach we proposed to address the data imbalance and neutral class performance issues is validated by recent literature. Onan and Balbal (2024) proved that ensemble data augmentation techniques, combining 'task-specific' and 'universal' transformations in Turkish text classification tasks, significantly improve model success and generalization capabilities. This finding provides strong evidence that the performance on minority classes (neutral and negative) observed in our study can be enhanced in future work through similar hybrid data augmentation methods.

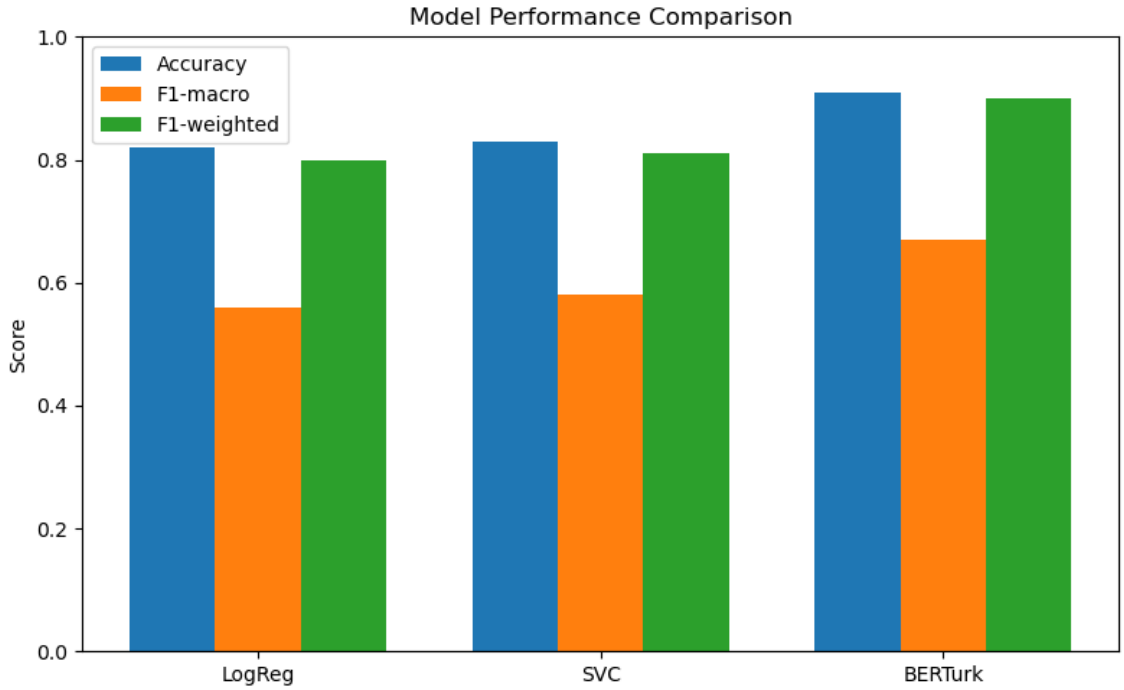

Figure 1. Comparison of model results.

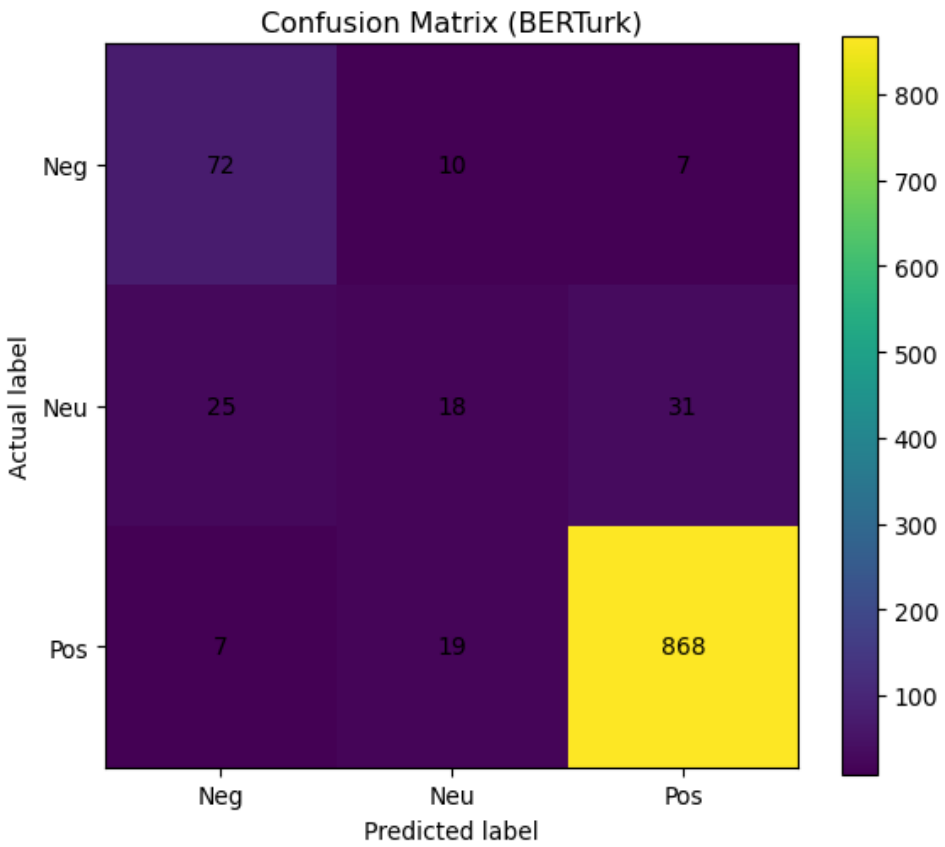

Figure 2. Confusion matrix (Berturk).

Despite these challenges, transformer models like BERTurk show promise in improving detection of minority classes. Çubukçu-Çerasi (2023), for instance, applied sentiment analysis to user comments on sustainable-consumption topics (including eco-friendly product reviews) and underscores the importance of capturing nuanced opinions in green domains. More directly, Incidelen and Aydoğan (2025) found that Turkish-specific transformers outperformed multilingual models in neutral review detection: their BERTurk variants achieved higher F1 on the neutral class than did XLM-RoBERTa or mBERT. They attribute this to the monolingual models' better grasp of Turkish grammar and context. In our work, the fine-tuned BERTurk model likely benefited similarly from deep contextual understanding, which helped it detect subtle sentiment cues that TF-IDF classifiers might miss. For example, phrases that could be interpreted as either neutral or mildly positive/negative are better disambiguated by BERTurk's contextual embeddings. This suggests that transformer-based methods can partly mitigate class imbalance effects by learning richer representations of minority-class examples.

In the specific context of "green" or sustainable product reviews, the findings have practical implications. Consumer attitudes toward eco-friendly products are of growing interest in business analytics. Özmen and Gündüz (2025) note that detailed sentiment analysis of Trendyol product feedback can help companies develop products that are both longer-lasting and more environmentally friendly, aligning with customer values. Likewise, Çubukçu-Çerasi's (2023) study on sustainable consumption comments highlights that understanding public sentiment around green products (through YouTube comments) can inform strategies for promoting sustainable consumption. Our study contributes to this emerging area by showing that advanced sentiment models (like BERTurk) can reliably extract positive and negative opinions in Turkish eco-product reviews, even if neutral attitudes are harder to capture. These insights can aid managers and marketers in Turkish e-commerce: for instance, high confidence in detecting positive/negative sentiment allows firms to identify and promote well-received green products, while being aware that ambiguous "neutral" feedback may need manual interpretation.

In summary, our results agree with prior literature in showing the clear benefit of transformer-based models over classical ML in Turkish review sentiment tasks. The drop in performance for neutral feedback reflects an ongoing issue noted by others. Compared to previous studies on Trendyol data, our overall accuracy and F1 scores are in the same range or higher. The gains on positive/negative sentiment detection with BERTurk demonstrate the value of deep learning for Turkish, as also reported elsewhere. Future work should continue to address neutral-class ambiguity (for example, via data augmentation or hierarchical modeling) and explore

more domain-specific green-product lexicons. Nevertheless, the convergence of our findings with existing research underscores the robustness of the conclusions: transformer-based methods excel in Turkish sentiment analysis, even as detecting minority neutral classes remains challenging.

5. Conclusion

In this study, we set out to evaluate the effectiveness of different sentiment classification models on Turkish-language e-commerce reviews, with a particular focus on products associated with environmental and sustainability themes. Among the three models tested, BERTurk consistently yielded the most accurate and balanced results. Its strong performance, especially in detecting clearly polarized sentiments (positive and negative), supports the growing consensus in recent literature that contextual language models offer tangible advantages over traditional machine learning methods in sentiment-related tasks involving morphologically rich languages like Turkish.

The traditional models we tested, while reliable in broader classification patterns, struggled to handle sentiment ambiguity and minority classes, especially in identifying neutral expressions. This is a common and well-documented issue in Turkish sentiment analysis, where the expression of neutrality tends to be less consistent, more culturally nuanced, and linguistically underrepresented. In our dataset, neutral comments were not only fewer in number but also more frequently misclassified—often mistaken for positive sentiment, which suggests that deeper linguistic understanding is needed to accurately resolve borderline or mixed opinions.

Our findings offer several practical takeaways for Management Information Systems and the wider field of e-commerce analytics. First, the ability of transformer-based models to interpret sentiment with higher fidelity can enhance customer feedback systems, particularly for platforms that promote green or sustainable products. As consumers increasingly express their expectations for environmental responsibility, accurate sentiment tracking can support strategic product development, sustainability positioning, and ethical brand communication. The use of robust sentiment classifiers also opens up new opportunities for automated monitoring of shifting public attitudes toward eco-friendly goods in real time.

Despite the success of the BERTurk model in our experiments, certain limitations remain. The imbalance in class distribution continues to skew performance metrics, especially those that rely on macro-averaging. Furthermore, while BERT-based models show impressive generalization, they still depend heavily on well-prepared and balanced training data. Future studies should consider constructing domain-adapted corpora that more accurately reflect the linguistic and emotional diversity of green consumer feedback. Additional research could also explore semi-supervised learning or ensemble approaches that explicitly target neutral sentiment or exploit complementary data sources like product metadata and user demographics.

To conclude, this study highlights the practical value of transformer-based language models in extracting actionable insights from Turkish-language sentiment data in e-commerce. The results confirm that BERTurk not only improves accuracy but also helps recover underrepresented sentiments more effectively than classical techniques. These advances can contribute meaningfully to MIS-driven decision-making processes, especially in understanding and shaping consumer attitudes toward sustainable products. Continued progress in this area will require both methodological innovation and richer, better-annotated Turkish datasets, ideally aligned with real-world commercial and social applications.

Author Contributions

The percentages of the author' contributions are presented below. The author reviewed and approved the final version of the manuscript.

	C.O.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C = concept, D = design, S = supervision, DCP = data collection and/or processing, DAI = data analysis and/or interpretation, L = literature search, W = writing, CR = critical review, SR = submission and revision, PM = project management, FA = funding acquisition.

Conflict of Interest

The author of this article declare that they have no conflict of interest

Ethical Consideration

Ethics committee approval was not required for this study because there was no study on animals or humans.

References

Aguilar-Moreno, J. A., Palos-Sánchez, P. R., & Pozo-Barajas, R. (2024). Sentiment analysis to support business decision-making: A bibliometric study. *AIMS Mathematics, 9*, 4337–4375. https://doi.org/10.3934/math.2024215

Bilik, M. (2023). Analyzing Challenges and Opportunities in the E-Commerce Industry of Turkey. *İzmir İktisat Dergisi, 38*(4),

1138-1151. https://doi.org/10.24988/ije.1262286

Brendel, A. B., Chasin, F., Mirbabaie, M., Riehle, D. M., & Harnischmacher, C. (2022). Review of design-oriented Green Information Systems research. *Sustainability, 14*(8), 4650. https://doi.org/10.3390/su14084650

Chiang, C.-T. (2024). A systematic literature network analysis of green information technology for sustainability: Toward smart and sustainable livelihoods. *Technological Forecasting and Social Change, 199*, 123053. https://doi.org/10.1016/j.techfore.2023.123053.

Çubukçu-Çerasi, C. (2023). Embracing green choices: Sentiment analysis of sustainable consumption. In *2023 International Conference on Research in Engineering, Technology and Science (ICRETS)* (pp. 254–255). https://dergipark.org.tr/en/download/article-file/3431062

Daza, A., González Rueda, N. D., Aguilar Sánchez, M. S., Robles Espíritu, W. F., & Chauca Quiñones, M. E. (2024). Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: A bibliometric analysis, systematic literature review, challenges and future works. *International Journal of Information Management Data Insights, 4*(2), 100267. https://doi.org/10.1016/j.jjimei.2024.100267

Demir, A. F. (2024). *Trendyol product comments* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/ahmetfurkandemr/trendyol-product-comments/

Doğan, A., & Kara, N. (2025). Sözcük Tabanlı Duygu Analizi: Sosyal Medya Paylaşımlarına Dayalı E-Ticaret Siteleri Memnuniyet Düzeyi Karşılaştırması. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 8(4), 1624-1643. https://doi.org/10.47495/okufbed.1604591

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

Huang, H., Zavareh, A.A. and Mustafa, M.B. (2023) Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions. *IEEE Access*, 11, 90367-90382. https://doi.org/10.1109/access.2023.3307308

Incidelen, M., & Aydoğan, M. (2025). Sentiment analysis in Turkish using language models: A comparative study. *European Journal of Technique, 15*(1), 68–74. https://doi.org/10.36222/ejt.1592448

Loke, R. E., & Pathak, S. (2023). Decision support system for corporate reputation based on social media sentiment analysis. In *Proceedings of the 18th International Conference on Evaluation of Novel Approaches to Software Engineering*. https://www.scitepress.org/Papers/2023/121364/121364.pdf

Maarif, M. R., Syafrudin, M., & Fitriyani, N. L. (2024). Uncovering Sustainability Insights from Amazon's Eco-Friendly Product Reviews for Design Optimization. *Sustainability*, *16*(1), 172. https://doi.org/10.3390/su16010172

Macías Urrego, J. A., García Pineda, V., & Montoya Restrepo, L. A. (2024). The power of social media in the decision-making of current and future professionals: a crucial analysis in the digital era. *Cogent Business & Management*, *11*(1). https://doi.org/10.1080/23311975.2024.2421411

Mangiaracina R, Marchet G, Perotti S, Tumino A (2015), "A review of the environmental implications of B2C e-commerce: a logistics perspective". *International Journal of Physical Distribution & Logistics Management*, Vol. 45 No. 6 pp. 565–591, doi: https://doi.org/10.1108/IJPDLM-06-2014-0133

Nave, M., Rita, P., & Guerreiro, J. (2018). A decision support system framework to track consumer sentiments in social media. *Journal of Hospitality Marketing & Management*, *27*(6), 693–710. https://doi.org/10.1080/19368623.2018.1435327

Öcal, A. (2025). BERT-based sentiment analysis of Turkish e-commerce reviews: Star ratings versus text. *Sakarya University Journal of Computer and Information Sciences, 8*(4), 677–687. https://doi.org/10.35377/saucis...1747068

Oláh, J., Popp, J., Khan, M.A., & Kitukutha, N. (2022). Sustainable e-commerce and environmental impact on consumer behaviour. *Economics and Sociology, 15*(2), 271–285. https://doi.org/10.14254/2071-789X.2023/16-1/6

Onan, A., & Balbal, K. F. (2024). Improving Turkish text sentiment classification through task-specific and universal transformations: An ensemble data augmentation approach. *IEEE Access*, 12, 4413–4458. https://doi.org/10.1109/ACCESS.2024.3349971

Özmen, C. G., & Gündüz, S. (2025). Comparison of machine learning models for sentiment analysis of big Turkish web-based data. *Applied Sciences, 15*(5), 2297. https://doi.org/10.3390/app15052297

Ramos, C. M. Q., Cardoso, P. J. S., Fernandes, H. C. L., & Rodrigues, J. M. F. (2023). A Decision-Support System to Analyse Customer Satisfaction Applied to a Tourism Transport Service. *Multimodal Technologies and Interaction*, *7*(1), 5. https://doi.org/10.3390/mti7010005

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. https://doi.org/10.1016/j.jksuci.2023.02.017

Shaik Vadla, M. K., Suresh, M. A., & Viswanathan, V. K. (2024). Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT. *Algorithms*, *17*(2), 59. https://doi.org/10.3390/a17020059

Singh N, Jung I, Han H, Ariza-Montes A, Vega-Muñoz A. (2022). Green Information System (GIS) Model in the Conference Sector: Exploring Attendees' Adoption Behaviors for Conference Apps. *Psychol Res Behav Manag*. 2022;15:2229-2243. https://doi.org/10.2147/PRBM.S370657

Teke, B., Yazıcı, S. N., Zamir, G., Budak, A. B., & Karabey Aksakallı, İ. (2025). BERTurk-based sentiment analysis on e-commerce multi domain product reviews. *Afyon Kocatepe University Journal of Science and Engineering, 25*(3), 497–509. https://doi.org/10.35414/akufemubid.1537513

UNCTAD. (2024). *E-commerce and environmental sustainability* (Chapter 5). United Nations Conference on Trade and Development. https://unctad.org/system/files/official-document/der2024_ch05_en.pdf

Wu, P., Tang, T., Zhou, L., & Martínez, L. (2024). A decision-support model through online reviews: Consumer preference analysis and product ranking. *Information Processing & Management, 61*(4), 103728. https://doi.org/10.1016/j.ipm.2024.103728

Zhang, Z., Guo, J., Zhang, H. Zhou, L., & Wang, M. (2022). Product selection based on sentiment analysis of online reviews: an intuitionistic fuzzy TODIM method. *Complex Intell. Syst.* **8**, 3349–3362 (2022). https://doi.org/10.1007/s40747-022-00678-w

Zümberoğlu, K. B., Dik, S. Z., Karadeniz, B. S., & Sahmoud, S. (2025). Towards better sentiment analysis in the Turkish language: Dataset improvements and model innovations. *Applied Sciences, 15*(4), 2062. https://doi.org/10.3390/app15042062