Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Understanding Aspect–Sentiment Drivers of Overall Ratings in Second-Hand Marketplace Apps through Text Analytics and Regression Analysis

Vicdan YALÇIN[1*], Ahmet Cumhur ÖZTÜRK[2], Mustafa ÇETİN[3]

[1] Aydın Adnan Menderes University, Management Information Systems Department, yalcinvicdan@hotmail.com, Orcid No: 0009-0005-6815-1942
[2] Aydın Adnan Menderes University, Database, Network Design and Management Department, cumhur.ozturk@adu.edu.tr, Orcid No 0000-0002-2677-3269
[3] Aydın Adnan Menderes University, Management Information Systems Department, mcetin@adu.edu.tr, Orcid No 0000-0001-8264-7657

ARTICLE INFO

Article history:

Received 25 September 2025
Received in revised form 15 November 2025
Accepted 23 November 2025
Available online 24 March 2026

Keywords:

Aspect-Based Sentiment Analysis, Ridge Regression, Customer-to-Customer (C2C) Platforms, Text Analytics, Online Ratings

Doi: 10.24012/dumf.1790105

* Corresponding author

ABSTRACT

User-generated reviews and star ratings strongly influence customer trust, download decisions, and platform reputation in mobile app marketplaces. This study analyzes reviews from three Turkish second-hand marketplace applications (Letgo, Dolap, Gardrops) on Google Play to uncover how specific aspect–sentiment patterns affect overall ratings. A multi-stage natural language processing (NLP) pipeline was applied. Review sentences were embedded with multilingual SBERT and clustered using KMeans, resulting in 13 higher-level aspect categories. Sentiment was classified using a domain specific Turkish ELECTRA model validated on 1,000 manually annotated sentences. Ridge regression was then employed to quantify the contribution of aspect–sentiment pairs to star ratings. The analysis showed that negative experiences related to returns, shipping costs and fraudulent practices consistently decreased ratings while positive mentions of usability, transaction satisfaction and customer support produced stronger rating improvements. Comparative findings revealed that higher-rated apps (Dolap: 4.4, Gardrops: 4.2) accumulated more positive experiences, whereas Letgo (2.7) exhibited recurring trust and fairness issues. These results highlight which service aspects most strongly shape customer evaluations. By linking aspect-level sentiment to rating outcomes, the study provides platform managers with actionable insights for improving satisfaction and strengthening competitiveness in second-hand consumer-to-consumer (C2C) marketplaces.

Introduction

User generated content particularly online reviews and star ratings in mobile app ecosystem are crucial for shaping customer trust, influencing download decisions and building platform reputation. An app's overall rating enhances customer engagement and acquisition, as it is widely interpreted as an indicator of reliability and service quality [1,2]. Although previous studies have explored how app ratings and review sentiment influence user behavior, fewer have examined how review content can be used strategically to improve an app's overall rating. This study investigates the relationship between review sentiment, aspect-level expressions and overall ratings across three different second-hand marketplace applications listed on the Google Play Store. While two of the apps have high overall ratings (4.2 and 4.4), one app has a significantly lower overall rating (2.7). By analyzing competing second-hand marketplace apps with stark rating differences, we identify not just sentiment patterns but which service aspects drive satisfaction or frustration.

To uncover the latent factors affecting overall app ratings, this study applied a multi-stage natural language processing (NLP) pipeline. First, sentence-level aspect extraction was performed by embedding review sentences with the multilingual SBERT model and clustering them using KMeans. The top TF-IDF terms from each cluster were then interpreted and merged into 13 higher-level aspect categories. Second, sentiment classification was conducted using a domain-specific Turkish ELECTRA-based classifier, which was validated against a manually annotated gold standard dataset of 1,000 review sentences. Finally, the resulting aspect–sentiment pairs were linked to overall star ratings through Ridge regression modeling. Ridge regression, a regularized form of linear regression, was chosen because it stabilizes coefficient estimates when features (aspect–sentiment pairs) are correlated. This modeling step allowed us to quantify which aspects contribute most strongly positively or negatively to overall app ratings.

Our methodology extends beyond descriptive analysis to offer a strategic roadmap for rating improvement, validated through comparative case studies. The analysis results provide actionable insights into how specific service issues influence overall ratings.

Based on this motivation, we propose the following research questions:

- RQ1: Which aspect–sentiment pairs most strongly influence user ratings in C2C platforms?
- RQ2: How do aspect–sentiment drivers differ between high-rated and low-rated C2C apps?
- RQ3: What practical strategies can be derived from these findings to improve overall app ratings?

For answering RQ1, we applied sentence-level regression analysis to identify the relative impact of different aspect–sentiment pairs on review ratings. Our findings demonstrate that negative experiences particularly returns, order problems, shipping costs and fraudulent or dishonest practices consistently depressed ratings. However, their negative effects were moderate in magnitude. In contrast, positive mentions of platform usability, transaction satisfaction, help requests and support inquiries contributed substantially larger gains. This indicates that while complaints reduce ratings, user satisfaction is more strongly driven by the presence of positive experiences. For answering RQ2 we compared Dolap, Gardrops and Letgo three C2C platforms that have clear differences in their overall ratings. The lower rated app (Letgo, 2.7) was characterized by a high density of unresolved issues related to trust, returns and platform fairness. In contrast, Gardrops (4.2) and Dolap (4.4) were more frequently associated with positive sentiments about usability, support and reliability. Although negative complaints were present across all platforms, the higher rated apps were better at accumulating positive experiences that outweighed these frictions, resulting in stronger overall scores. For answering RQ3, we interpreted the regression outcomes as actionable insights for platform managers.

Literature Review

The overall star rating of a mobile app serves as a critical indicator of app quality, user trust and marketplace competitiveness. The high rating of a mobile app increases downloads and higher revenue [3, 4]. While numerical ratings provided by users gives a quick summary of user experiences, the textual content of reviews makes it possible to understand deep insights about user concerns and satisfaction drivers [5]. An increasing number of studies have investigated the relationship between review sentiments and star ratings.

A growing body of research demonstrates how sentiment analysis and aspect-based sentiment analysis (ABSA) can explain or predict user ratings. For example, sentiment analysis was applied to mobile banking app reviews in the Canadian market to extract drivers of satisfaction and dissatisfaction [4]. It has been demonstrated how different aspects influence customer perceptions and satisfaction in e-commerce reviews using aspect-level sentiment analysis [6]. Similarly, sentiment strength analysis has been utilized for review rating prediction, and systematic reviews of sentiment analysis in social media have reinforced its value in understanding customer satisfaction [7, 8].

Recent advances in ABSA have leveraged transformer models, often in combination with structural enhancements. For instance, a framework was introduced that integrates BERT with multi-layered graph convolutional networks, enabling models to capture fine-grained word relationships and outperform previous approaches [9]. In another study, Latent Dirichlet Allocation (LDA) topic modeling with a TF-BERT classifier, showing robust cross-domain performance [10]. A systematic review emphasized that, despite rapid methodological advances, many ABSA studies suffer from limited domain diversity particularly regarding peer-to-peer platforms and non-English contexts highlighting a critical gap [11]. In the context of Turkish language processing, low-rated Turkish app reviews were analyzed using SBERT-based semantic similarity modeling, demonstrating that complaint clustering can be enhanced through sentence-level analysis [12].

Empirical findings also confirm that aspect sentiments are highly predictive of overall ratings. Accuracy levels exceeding 90% have been reported in aspect sentiment classification across 14 purchase-process categories using RoBERTa, demonstrating the predictive strength of aspect polarity [13]. Similarly, it has been shown that aggregating aspect-level sentiment intensities yields strong predictive power for overall star ratings [5]. These findings establish a robust theoretical foundation for aspect-aware modeling in rating prediction.

Linear regression has also been employed to model the relationship between review text and ratings. By combining ABSA with regression to predict review ratings in e-commerce, demonstrating that aspect sentiment signals improve predictive accuracy [14]. Regression has been applied to sentiment scores and integrated with topic features in hotel reviews to enhance rating prediction. However, these studies primarily treated regression as a predictive tool and did not leverage regression coefficients to analyze which specific aspect sentiments drive ratings upward or downward. To the best of our knowledge, no prior research has combined large language model embeddings with regression analysis to identify both the direction (positive or negative) and magnitude of aspect-level effects on app ratings [7, 15].

In sentiment–rating modeling, only a few studies have examined peer-to-peer second-hand marketplaces, whereas most research has focused on hospitality, restaurants, and general e-commerce [6]. Yet, second-hand marketplaces pose unique challenges such as trust, fraud prevention, transaction reliability, and return policies [16, 17]. In addition, applications in non-English contexts remain underrepresented, particularly in morphologically rich languages like Turkish, where preprocessing complexities further complicate ABSA. Also big data research indicates that large scale user-generated content introduces additional challenges related to data privacy, security and scalable processing architectures, highlighting the need for robust analytical frameworks in high volume mobile commerce environments [18, 19].

The present study addresses these gaps by focusing on second-hand marketplace applications in Turkey, an underexplored yet rapidly growing segment of mobile commerce. We integrate Turkish-specific preprocessing (via Zemberek) with SBERT-based embeddings to conduct ABSA and apply linear regression not only for rating prediction but also as a diagnostic tool to determine which aspect sentiments drive ratings upward or downward across competing platforms. By comparing applications with contrasting aggregate ratings, this study contributes both theoretical insights and actionable recommendations for improving user experience and platform performance.

Methodology

The methodological workflow of this study is shown in Figure 1 and consists of six main stages: data collection, preprocessing, vectorization, topic modeling, sentiment detection and evaluation. In the first stage a review dataset was generated by collecting user review texts, date of each review and rating of each review from Google Play. Then preprocessing was applied to raw text including steps such as language filtering, emoji and number removal, lowercasing, sentence segmentation, lemmatization and stopword removal for standardizing. In the third step, review sentences were transformed into dense numerical representations using SBERT embeddings. Next these embeddings were then grouped into coarse themes through clustering based topic modeling. Following this, the polarity of each review sentence was detected with using a transformer based model. Finally, a regression analysis was employed for evaluating the influence of aspect level sentiments on overall review ratings.

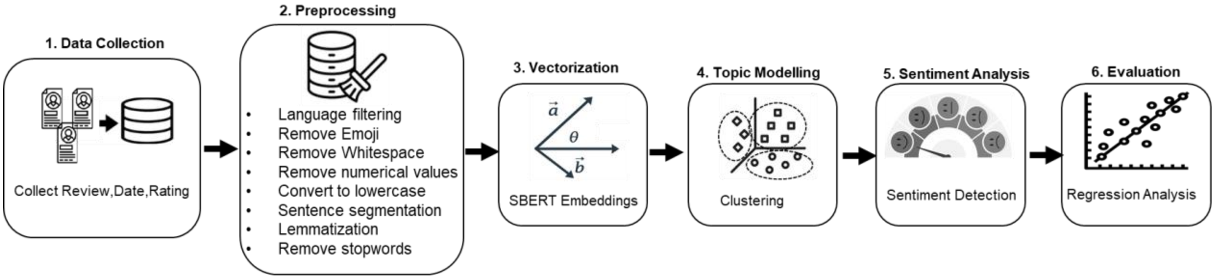

Figure 1. Methodological workflow of our study.

Data Collection and Preprocessing

The dataset used in this study was collected from Google Play Store app reviews covering the period 2023 and 2024 of three leading consumer to consumer (C2C) marketplace applications in Turkey; Dolap, Letgo and Gardrops. Reviews were gathered manually to ensure accuracy and completeness. The reviews were collected manually through the Google Play web interface by navigating to each application's review section and iteratively loading all available comments. For each review, the text, posting date, and user-provided rating were copied into a structured dataset. Manual collection ensured the inclusion of all visible reviews without relying on API restrictions or automated scraping tools, which often limit access or truncate older comments.

Each review has its rating score between 1 and 5 and date of posting. A total of 49,592 reviews were collected where 17,628 of them belonging to Dolap, 22,702 of them belonging to Letgo and 9,262 of them belonging to Gardrops. After the data is collected we applied a multi step preprocessing pipeline for improving the performance of machine learning and data analysis approaches used in this study. First, non-Turkish reviews and very short reviews containing fewer than three words were removed from the dataset, as such entries do not provide sufficient information for meaningful Turkish aspect-based sentiment analysis. The language detection was performed using the langdetect library, that resulted in the elimination of 5,119 non-Turkish and short reviews. In the next step, non-textual elements; emojis, excessive whitespace, numerical-only tokens and irrelevant special characters were removed. All text was then normalized to lowercase. The reviews were subsequently segmented into individual sentences using the spaCy multilingual model, yielding a total of 74,320 sentences. Sentence segmentation was applied to ensure that aspect sentiment associations were captured at the sentence level. Following this, lemmatization was performed with Zemberek [20] to reduce words to their canonical forms, and stopwords were removed using a predefined Turkish stopword set [21]. The resulting cleaned and lemmatized sentences were then prepared for the next stage, where they were converted into dense vector representations using SBERT embeddings, which served as the basis for clustering and sentiment classification. Sentence-level SBERT embeddings were combined to create review-level representations by taking their weighted average. Sentences that mentioned more aspects were assigned higher weights so that aspect-relevant information had a stronger influence on the final review vector.

Table 1 presents a sample of the preprocessed dataset used in this study. The ReviewID column indicates the unique id of review that each review sentence belongs. The same review id may appear in more than one row as multi sentence reviews are split into separate rows. The Date column indicates when the review was posted. The Review Sentence column contains the cleaned and normalized text. Finally, the Rating column shows the numerical rating for the entire review provided by the user which is between 1 and 5.

Table 1. Representative sample of review sentences with ratings after preprocessing.

ReviewID	Date	Review Sentence	Rating	App
0	2024-12-31	(TR[a]) gerçekten şaka gibi saçma sapan bir sebepten yıllardır kullandığım uygulamaya giremiyorum (EN[b])i cannot access the app anymore for a ridiculous reason even though i have been using it for years	1	Dolap
0	2024-12-31	(TR[a])hesabıma tekrardan girmek istiyorum güvenlik ile ilgili hiçbir kuralı ihlal etmedim (EN[b])i want to log back into my account I did not violate any security rule	1	Dolap
13358	2024-05-15	(TR[a])geçen hafta gönderilen ürün dün teslim edildi (EN[b])the product sent last week was delivered yesterday	3	Gardrops
13358	2024-05-15	(TR[a])bir de saat ödeme onayi bekliyoruz (EN[b])we are waiting for payment approval for hours	3	Gardrops
13358	2024-05-15	(TR[a])onaydan sonra da saat paramızı çekmeyi beklicez anliyacaginiz günde (EN[b])after approval, we also wait hours to withdraw our money meaning we only get paid the next	3	Gardrops
25380	2024-12-28	(TR[a])son güncellemeden sonra ilan verilmiyor sürekli hata verip kapanıyor gereksiz bir uygulamaya dönüştü (EN[b])after the last update listings cannot be posted it constantly gives errors and turned into a useless app	1	Letgo

[a]Turkish,[b]English

Note: User-generated spelling errors were intentionally retained to preserve the authenticity of the original review texts.

Topic Modeling and Aspect Extraction

Topic modeling is the process of grouping semantically related sentences or documents into coherent themes, thereby enabling the identification of latent aspects within large collections of textual data. In this study, topic modeling was carried out using a clustering-based approach. Each review sentence was first transformed into a dense semantic vector representation with the multilingual SBERT model. These embeddings were then partitioned into clusters using the KMeans algorithm with k=50, which iteratively minimizes within-cluster variance while maximizing separation between clusters. To justify this choice, we experimented with multiple k values (20, 30, 40, 50, 60) and manually examined the semantic coherence of clusters. Among these alternatives, k=50 produced the most balanced structure without over-fragmentation or overly broad clusters. To interpret and label the resulting clusters, the top 10 representative words for each cluster were extracted based on their TF-IDF weights. Clusters that shared overlapping or highly similar top TF-IDF terms were subsequently merged, yielding 13 higher-level aspect dimensions. This merging step allowed conceptually similar fine-grained clusters to be grouped under broader, coherent aspect categories. Aspect names were assigned to these final clusters according to their most representative TF-IDF terms. In this way, the analysis produced 13 coherent and interpretable aspects that capture the major themes of user reviews. These aspect categories served as the foundation for the subsequent sentiment detection and regression analysis. The final aspect names, definitions and number of review sentence each aspect cluster contains are presented in Table 2.

The count column in Table 2 displays the distribution of review sentences across the 13 final aspect dimensions. As can be seen from the Table Shipping Costs, Updates and Feature Adjustments and Overall Platform Opinion emerge as the most frequently discussed themes, indicating that users place significant emphasis on transaction related costs, system changes and general satisfaction or dissatisfaction with the platform. In contrast, aspects such as Returns and Fees appear less frequently, suggesting that while they are less commonly mentioned, they may still carry important implications for user experience.

Table 2. Aspect dimensions identified after merging clusters.

Cluster ID	Aspect Name	Definition	Count
1	Shipping Costs	User concerns about delivery charges, free shipping policies and cost of logistics.	7,728
2	Updates, Version Changes & Feature Adjustments	Complaints or feedback about app updates, interface changes, or feature modifications.	5,737
3	Overall Platform Opinion	General evaluations of the platform, including overall satisfaction or dissatisfaction.	5,716
4	App Reliability & Performance	Issues with app crashes, errors, speed and technical malfunctions.	4,947
5	Platform Usability & Transaction Satisfaction	Experiences with ease of navigation, listing, searching, and satisfaction with transactions.	4,858
6	Authentication & Account Management	Problems with login, verification, password, or account access.	4,785
7	Orders	Complaints and experiences related to order placement, tracking and delivery.	2,429
8	Messaging & Notification Issues	Problems with in-app messaging, communication, and notifications.	2,016
9	Commissions	Concerns about seller/buyer commission rates and deductions.	1,539
10	Help Requests & Support Inquiries	User appeals for assistance, support requests, and customer service expectations.	1,114
11	Fraud, Scam & Dishonest Practices	Reports of fraudulent behavior, scams, or dishonest practices on the platform.	994
12	Returns	Experiences related to returning products, refunds, or exchanges.	695
13	Fees	Experiences about service fees, protection fees or transaction related charges (excluding shipping/commissions).	372

Sentence Level Sentiment Labelling and Validation

Sentiment analysis is a NLP task and it is the process of finding the emotional tone of a given text as positive, negative or neutral. It helps to capture the attitudes and emotions of customers when it is applied to online reviews. Traditional sentiment analysis techniques, such as lexicon based approaches or bag-of-words representations, often fail to capture contextual nuances, idiomatic expressions and sarcasm which are mostly appear in human generated texts. In contrast, transformer based pretrained language models can recognize these contextual nuances, idiomatic expressions and sarcasm. The Hugging Face platform provides an open repository of state-of-the-art pretrained NLP models. Because these models are trained on large scale corpora, they enable high performance sentiment classification. In this study, we employed a domain specific Turkish ELECTRA based classifier, incidelen/electra-base-turkish-sentiment-analysis-cased [22], a transformer model available through the Hugging Face platform, to classify the sentiments of review sentences. This classifier was fine-tuned on the TRSAv1 e-commerce reviews corpus, which contains 150,000 reviews in total, evenly distributed across three classes: 50,000 positive, 50,000 negative, and 50,000 neutral. We also evaluated the robustness of the sentiment analysis results using the classifier MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 [23], which is a multilingual zero-shot model based on DeBERTa-v3. To decide which model to use in sentiment analysis we created a gold standard dataset through a manually labeled gold standard dataset using stratified proportional random sampling to ensure balanced representation across platforms and aspects. Quotas were allocated proportionally based on platform size and within each platform, samples were distributed across aspects according to their frequency. Using this procedure, a total of 1,000 sentences were manually annotated as Positive, Negative, or Neutral for model comparison. The annotation process was conducted by 2 independent researchers and for ensuring reliability, inter-annotator agreement was calculated using Cohen's Kappa coefficient, which resulted in a score of 0.85, indicating strong agreement.

The comparative performance evaluation of Turkish ELECTRA and DeBERTa models is shown in Table 3. In the performance evaluation, accuracy and class-level F1-scores are used, where F1 is defined as the harmonic mean of precision and recall [24]. The Turkish ELECTRA model (incidelen/electra-base-turkish-sentiment-analysis-cased) demonstrated strong performance, achieving an overall accuracy of 0.949. At the class level, it identified negative sentences with an F1-score of 0.977, positive sentences with an F1-score of 0.896, while its performance on the neutral class was lower at 0.370. In comparison, the DeBERTa-based zero-shot model (MoritzLaurer/ mDeBERTa-v3-base-xnli-multilingual-nli-2mil7) obtained lower overall performance, with an accuracy of 0.847. Its class-level F1-

scores were 0.912 for negative sentences, 0.680 for positive sentences, and 0.114 for neutral sentences.

Table 3. Performance comparison of evaluated sentiment classification models.

Model	Accuracy	F1| Negative	F1|Positive	F1|Neutral
Turkish ELECTRA	0.949	0.977	0.896	0.370
DeBERTa	0.847	0.912	0.680	0.114

As the Turkish ELECTRA model achieved a higher overall accuracy and performance on the negative and positive classes, it was selected as the sentiment classifier for the subsequent analyses.

Since ratings of reviews are not in sentence level, before the regression analysis sentence level aspect–sentiment labels were aggregated into a single feature vector per review. For each review, all sentences belonging to the same review ID were grouped and binary indicators were created for each aspect–sentiment pair. A feature was assigned the value 1 if at least one sentence in the review mentioned that aspect with a given polarity and 0 otherwise. This binary aggregation prevents reviews with repeated mentions of the same aspect–polarity pair from dominating the feature space, while still capturing whether that signal is present at least once in the review. Sentence level sentiment scores were not averaged as the regression model aims to capture whether an aspect sentiment is expressed at least once in a review, which aligns with established ABSA feature engineering approaches [25].

Modeling the Relationship Between Aspect Sentiments and Review Ratings

To examine whether the review ratings in our datasets align with the overall app ratings reported in the Google Play Store, we compared the average review ratings of the three platforms. The mean values of review ratings were calculated as 2.71 for Gardrops, 2.31 for Dolap, and 1.59 for Letgo, which are broadly consistent with the corresponding official app ratings of 4.2, 4.4, and 2.7. We calculated Pearson correlation coefficient r between the dataset averages and Google Play ratings to formally assess this relationship and we obtained r≈0.99. The Pearson coefficient ranges from –1 to +1 and it indicates the strength and direction of a linear relationship between two variables value. The near perfect positive value (r≈0.99) demonstrates that higher average review ratings in the dataset are strongly associated with higher overall app ratings in the store, confirming the representativeness of our collected data. As a minor limitation, the average ratings in our manually collected dataset differ slightly from the official Google Play Store scores, likely due to time-dependent changes in user activity and the way Google Play aggregates ratings. The strong correlation indicates that the dataset remains adequately representative for analysis.

Based on this validation, we next sought to quantify the impact of aspect-level sentiments on review ratings using a linear regression model. Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It is particularly useful in behavioral and text-mining studies because it provides interpretable coefficients that capture both the direction (positive or negative) and strength of associations. A multiple linear regression model for predicting review ratings can be expressed as follows [26]:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_i + \epsilon_i \qquad (1)$$

In equation (1) y_i denotes review rating for ith review. β_0 is the is the baseline rating, it's the rating of a review would get if it had none of the features under consideration. The baseline rating is calculated as the mean of all review ratings adjusted for the average contribution of each feature. It reflects the overall average rating in the dataset after accounting for how often features appear on average. β_j captures the impact of a specific feature which in our study corresponds to an aspect. A positive β_j indicates that the feature is a rating booster whereas a negative β_j indicates that the feature is a rating reducer. x_i represents the presence or absence of a feature where in our case it is 1 or 0. Finally ϵ_i is the error term that accounts for unexplained variation.

In the context of our study we considered each aspect and its corresponding sentiment polarity as feature such as “Orders & Returns| Positive” or “Cross-App Comparisons| Negative”. For each review, these features are encoded as binary for indicators (present or not) and then the review rating is regressed on these features. For example, if the coefficient β_j “Orders & Returns | Positive” is +0.45, this means that reviews mentioning “Orders & Returns” in a positive way are predicted to have ratings about 0.45 points higher than reviews without that feature while holding all other features constant. On the other hand, if the coefficient for “Cross-App Comparisons | Negative” is −0.60 this means that reviews having this aspect would reduce the expected rating by 0.6 points while holding all other features constant. The regression analysis results directly reveal which aspects and sentiments exert the strongest positive or negative influence on user evaluations.

Table 4. Example of binary encoding of aspect–sentiment features at the review level.

Review_Id	Orders & Returns | Positive	Cross-App Comparisons | Positive	Cross-App Comparisons | Negative	Review Rating
1	1	1	0	5
2	0	0	1	3

For illustration, suppose we have two reviews (with IDs 1 and 2) and three aspect sentiment features, as shown in Table 4 In this table, Review with id 1 consists of two sentences where one mentioning Orders & Returns with a positive sentiment and another mentioning Cross-App Comparisons with a positive sentiment. Review with id 2 contains a single sentence that mentions Cross-App Comparisons with a negative sentiment. Using this table, the regression model estimates coefficients for each feature together with the intercept. Suppose the model yields

β_0=+4.0, β_1=+0.8, for Orders & Returns | Positive, β_2=+0.5 for Cross-App Comparisons | Positive, and β_3=−1.0 for Cross-App Comparisons | Negative. Based on these estimates, the predicted rating for Review 1 is y_1= 4.0+(0.8)*1+(0.5)*1=5.3 which is close to the corresponding review's rating 5. For Review 2, the predicted review rating is y_2=4.0-(1.0)*1=3.0 exactly matching the review's rating.

This simplified example illustrates how we encoded aspect–sentiment features and estimated their contribution to ratings. In the actual analysis, however, the feature space was much larger and many aspect–sentiment variables were correlated. Ordinary multiple linear regression can become unstable under such conditions, producing inflated or unreliable coefficients. To overcome this issue, we employed Ridge regression, a regularized extension of linear regression. Ridge regression adds an *L2* penalty term that shrinks coefficient estimates toward zero, reducing variance while retaining interpretability. The Ridge estimator minimizes the following objective function [26]:

$$\hat{\beta} = \underset{\beta}{argmin} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (2)$$

In equation (2) y_i is the rating of the ith review sentence, β_0 is the starting value of the model called intercept, β_j is the effect of each aspect x_{ij} and λ is the parameter that controls regularization. Ridge regression works like linear regression but includes a penalty that shrinks the coefficients toward zero. The strength of this shrinkage is determined by λ larger values of λ produce stronger shrinkage. This regularization makes the model more stable and reduces overfitting when there are many or correlated aspects, as in our sentence-level aspect–sentiment data. The regularization parameter λ was selected through 10-fold cross-validated grid search over the range {0.01, 0.1, 1, 10, 100}. The optimal value was λ=1.0, which yielded the lowest validation error. For illustration, returning to the example in Table 4, ordinary regression produced coefficients of β_1=+0.8, β_1=+0.8 and β_3=−1.0. With Ridge regression, these estimates are slightly reduced in magnitude, for example to β_1=+0.7, β_2=+0.4 and β_3=−0.9, while the intercept remains at β_0=4.0. The predicted ratings for the two reviews remain close to 5 and 3, but the coefficients are now more stable and less sensitive to correlations among aspects. In this way, Ridge regression changes the estimates by keeping predictions similar while improving robustness.

Results

Sentence Level Aspect Distributions Across Sentiments

In Figure 2 the distribution of review aspects across the three sentiment categories of positive, negative and neutral is shown for the platforms Dolap, Gardrops and Letgo. Each bar in Figure 2 represents the proportion of different aspects across sentiment categories within the review set of each platform. The colored bar chart format provides a comparative view of how specific topics are emphasized differently across platforms and sentiments. Taller segments within a bar indicate a higher proportion of sentences belonging to that topic. This bar chart visualization makes it possible to identify the dominant themes in positive, negative and neutral feedback while also highlighting cross platform similarities and differences in terms of user concerns.

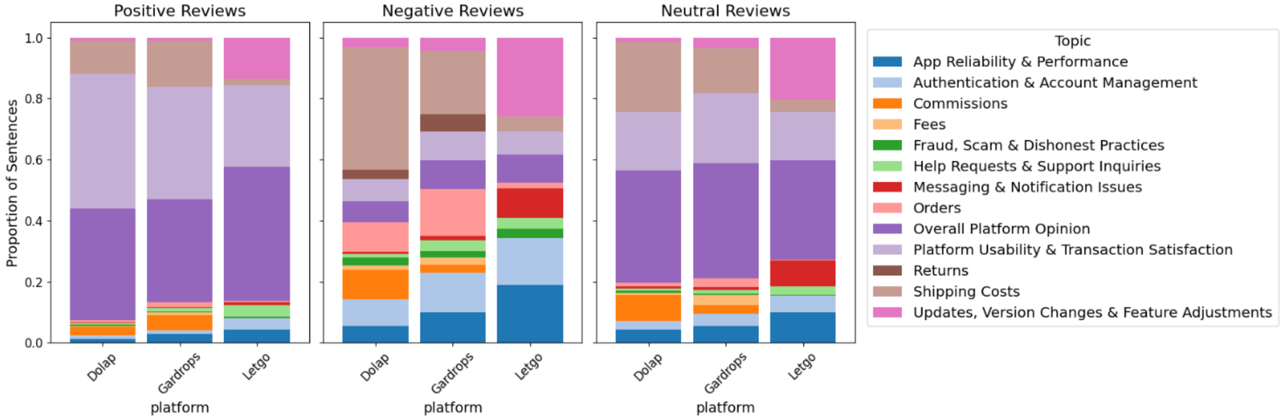

Figure 2. Sentiment distribution of review sentences across topics.

According to figure 2 negative reviews show a more diverse topic distribution compared to positive and neutral reviews in each platform. For Dolap, negative review sentences are spread across several issues, with the largest share related to Shipping Costs followed by Orders, Commissions and App Reliability and Performance. Gardrops displays a similar structure to Dolap where Orders and Authentication and Account Management also stand out as major concerns. In contrast, Letgo differs noticeably, as negative review sentences are dominated by Updates, Version Changes and Feature Adjustments together with App Reliability & Performance, while Fraud, Scam & Dishonest Practices and Authentication & Account Management also appear prominently. This divergence in the distribution of aspects on Letgo reflects its position as the app with the lowest overall rating. In contrast, Dolap and

Gardrops show a more balanced distribution of negative topics, which correlates with their higher average ratings.

Neutral reviews across platforms are largely shaped by *Overall Platform Opinion and Platform Usability & Transaction Satisfaction*. While Dolap and Gardrops follow this structure, Letgo differs by showing higher proportions of *Updates, Version Changes & Feature Adjustments, Orders and App Reliability and Performance.*

Positive review sentences are generally less fragmented and dominated by a few recurring aspects across platforms. In Dolap and Gardrops, the most frequent positive mentions relate to *Overall Platform Opinion and Platform Usability & Transaction Satisfaction* which highlights user appreciation of the general shopping experience. Letgo shows a similar trend, although with a stronger emphasis on *Updates, Version Changes & Feature Adjustments* compared to the other two platforms. This shared focus on platform usability and general opinion in Dolap and Gardrops again aligns with their higher ratings, whereas Letgo's emphasis on system related updates reflects a different profile of user satisfaction.

In summary, Figure 2 shows clear differences in how review sentence aspects are distributed across sentiments and platforms. Positive reviews are dominated by *Overall Platform Opinion and Platform Usability & Transaction Satisfaction* on all three platforms, with Letgo placing additional emphasis on *Updates, Version Changes & Feature Adjustments*. On the other hand, negative reviews present the most diverse distribution, with Dolap and Gardrops showing higher proportions of *Shipping Costs, Orders and Commissions, while Letgo stands out with more focus on Updates, Version Changes & Feature Adjustments, App Reliability & Performance and Fraud, Scam & Dishonest Practices.* Neutral reviews sentences concentrated around *Overall Platform Opinion and Platform Usability & Transaction Satisfaction*, although Letgo differs from Dolap and Gardrops by placing more weight on *Updates, Version Changes & Feature Adjustments, Orders, App Reliability & Performance.*

Impact of Aspect Sentiment Pairs on Ratings based on Regression Analysis

Figure 3 presents the estimated effects of all aspect–sentiment pairs that occurred in at least 10 reviews per platform, with coefficients obtained from Ridge regression models fitted separately for Dolap, Gardrops, and Letgo. The minimum frequency threshold (≥10) was applied to exclude very rare mentions that could yield unstable estimates. Three rare pairs, *Fraud, Scam & Dishonest Practices | Positive, Returns | Positive and Messaging & Notification Issues | Positive* were excluded to prevent unstable coefficient estimates. As a result, Figure 3 displays 23 pairs, which represent all aspect–sentiment combinations with sufficient support in the data.

Ridge regression was employed to address multicollinearity among correlated aspects and to stabilize coefficient estimates, while still retaining interpretability. The y-axis lists the aspect–sentiment pairs, while the x-axis represents the three platforms. The color scale shows whether an aspect–sentiment pair increases or decreases ratings and by how much. Green indicates that the aspect is associated with higher ratings (a positive effect), while orange indicates that the aspect is associated with lower ratings (a negative effect). Darker shades represent stronger effects, whereas lighter shades represent weaker effects.

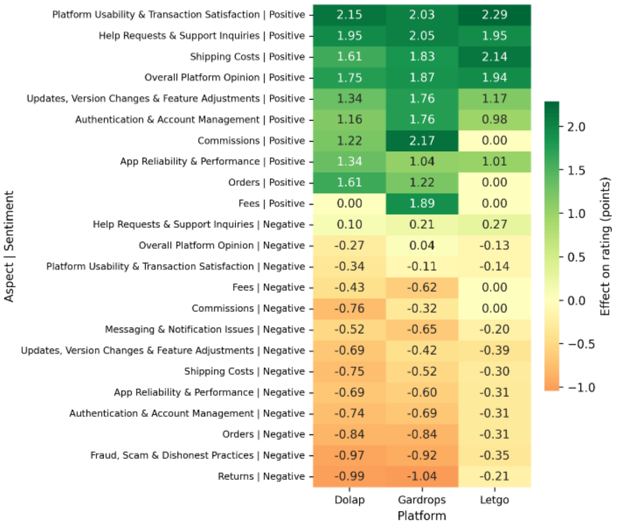

Figure 3. Top drivers of ratings across platforms (Ridge coefficients).

In Figure 3 there is a clear asymmetry is observed between negative and positive mentions. Negative mentions of *Returns*, Fraud and Scam & Dishonest Practices, *Orders*, *Authentication & Account Management*, *App Reliability & Performance* and *Shipping Costs* consistently reduce ratings on all three platforms, with coefficient magnitudes ranging from –0.2 to –1.0 points. By contrast, positive mentions exert substantially stronger upward effects. *Platform Usability & Transaction Satisfaction* emerges as the strongest driver, increasing ratings by approximately +2.15 to +2.29 points across platforms. This is followed by *Help Requests & Support Inquiries* (+1.95 to +2.05), *Shipping Costs* (+1.61 to +2.14) and *Overall Platform Opinion* (+1.75 to +1.94). Additional but smaller gains are also observed for *Updates, Version Changes & Feature Adjustments* and *Authentication & Account Management* when expressed positively.

The Figure 3 also shows that negative mentions of aspects such as Returns, Orders, Shipping Costs and Fraud, Scam & Dishonest Practices lower ratings, but their negative impact is relatively small. By contrast, positive mentions of Platform Usability & Transaction Satisfaction, Help Requests & Support Inquiries and Overall Platform Opinion increase ratings much more strongly. This means that overall ratings are shaped more by positive aspect sentiment pairs than by negative ones.

For C2C platforms, these findings suggest that overall ratings can be most effectively improved by reinforcing ease of use, ensuring reliable performance and providing timely customer support, while continuing to minimize recurring transaction and trust issues. By systematically strengthening these positive drivers, platforms can enhance user satisfaction and achieve higher ratings in competitive digital marketplaces.

Discussion

Our findings show that aspect-level sentiment signals are strong predictors of overall star ratings in second-hand marketplace applications. This is consistent with prior research demonstrating that sentiment polarity and aspect-level cues strongly influence user ratings across various online platforms. Similar to Wang et al. [5] and Zhang et al. [7], we observed that aggregated sentiment intensities correlate closely with rating outcomes. However, our approach extends previous studies by using regression coefficients as diagnostic indicators, revealing which specific aspects drive ratings upward or downward across competing platforms. This interpretability offers a unique contribution to ABSA research and helps explain satisfaction dynamics in under-studied Turkish second-hand marketplaces.

Conclusion

The overall rating of a mobile app is strongly correlated with individual review ratings. Building on this knowledge, our study identified which aspect sentiment pairs most strongly influence user ratings in C2C platforms. This made it possible to determine the areas that should be prioritized to raise overall platform ratings. The analysis was conducted on review data collected from three widely used C2C applications in Turkey Dolap, Gardrops and Letgo.

The results demonstrate that while negative experiences such as Returns and Orders, Shipping Costs and Fraud, Scams & Dishonest Practices lower overall ratings, their effects are moderate compared to the much larger gains associated with positive mentions of Platform Usability & Transaction Satisfaction, Help Requests & Support Inquiries, and Overall Platform Opinion. This brings out that overall rating of a mobile app is more strongly shaped by positive experiences than by negative complaints.

Our findings indicate that C2C platforms can raise their overall ratings most effectively by reinforcing ease of use, ensuring reliability, and providing responsive customer support. Strengthening these positive drivers would have a more effective path toward higher ratings than focusing solely on eliminating negative issues. In this context, positive reviews offer valuable insights about the experiences contributing to high app ratings, while negative reviews reveal the underlying issues that may lower overall ratings. Understanding these relationships enables targeted improvements aimed at raising an app's overall score.

Beyond these practical implications, the study contributes methodologically by showing how sentence-level aspect–sentiment analysis combined with Ridge regression yields interpretable results in large-scale review mining. This approach extends beyond descriptive sentiment analysis to quantify which service features drive ratings, offering a replicable framework for other digital platforms and review-based ecosystems. The originality of this work lies in bridging advanced text analytics with interpretable statistical modeling to uncover rating drivers at the aspect sentiment level in C2C marketplaces, a domain that has received limited attention in prior information science research.

References

[1] M. Harman, Y. Jia, and Y. Zhang, "App store mining and analysis: MSR for app stores," in Proc. 9th IEEE Working Conf. Mining Softw. Repositories (MSR'12), 2012, pp. 108–. doi: 10.1145/2804345.2804346.

[2] H. Sällberg, S. Wang, and E. Numminen, "The combinatory role of online ratings and reviews in mobile app downloads: An empirical investigation of gaming and productivity apps from their initial app store launch," J. Marketing Analytics, vol. 11, no. 3, pp. 426–442, 2022. doi: 10.1057/s41270-022-00171-w.

[3] S. Ba, S. He, and S. Lee, "Mobile app adoption and its differential impact on consumer shopping behavior," Prod. Oper. Manag., vol. 31, no. 2, pp. 764–780, 2022. doi: 10.2139/ssrn.3727035.

[4] Y. Amirkhalili and H. Y. Wong, "Banking on feedback: Text analysis of mobile banking iOS and Google app reviews," arXiv Preprint, arXiv:2503.11861, 2025. doi: 10.48550/arXiv.2503.11861.

[5] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdiscip. Rev. Data Mining Knowl. Discov., vol. 10, no. 3, e1331, 2020. doi: 10.1002/widm.1253.

[6] S. Vanaja and M. Belwal, "Aspect-level sentiment analysis on e-commerce data," in 2018 Int. Conf. Inventive Res. Comput. Appl. (ICIRCA), 2018, pp. 1275–1279. doi: 10.1109/ICIRCA.2018.8597286.

[7] Y. Wang, Y. Huang, and M. Wang, "Aspect-based rating prediction on reviews using sentiment strength analysis," in Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst., Springer, 2017, pp. 439–447. doi: 10.1007/978-3-319-60045-1_45.

[8] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," Procedia Comput. Sci., vol. 161, pp. 707–714, 2019. doi: 10.1016/j.procs.2019.11.174.

[9] K. Aziz, D. Ji, P. Chakrabarti, T. Chakrabarti, M. S. Iqbal, and R. Abbasi, "Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection," Sci. Rep., vol. 14, no. 1, 14646, 2024. doi: 10.1038/s41598-024-61886-7.

[10] R. J. Dhanal and V. R. Ghorpade, "Aspect-based sentiment analysis using topic modelling and machine learning," Int. J. Electr. Comput. Eng., vol. 14, no. 6, 2024. doi: 10.11591/ijece.v14i6.pp6689-6698.

[11] Y. C. Hua, P. Denny, J. Wicker, and K. Taskova, "A systematic review of aspect-based sentiment analysis: Domains, methods, and trends," Artif. Intell. Rev., vol.

57, no. 11, 296, 2024. doi: 10.1007/s10462-024-10906-z.

[12] A. C. Öztürk, "Large Group Decision Making for Aspect-Level Consensus Evaluation in Low-Rated App Reviews", Müh.Bil.ve Araş.Dergisi, c. 7, sy. 2, ss. 173–184, 2025, doi: 10.46387/bjesr.1716998.

[13] L. Davoodi, J. Mezei, and M. Heikkilä, "Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms," Electron. Commer. Res., pp. 1–43, 2025. doi: 10.1007/s10660-025-09948-4.

[14] S. Gojali and M. L. Khodra, "Aspect-based sentiment analysis for review rating prediction," in 2016 Int. Conf. Adv. Informatics: Concepts, Theory Appl. (ICAICTA), 2016, pp. 1–6. doi: 10.1109/ICAICTA.2016.7803110.

[15] Y. Putranto, B. Sartono, and A. Djuraidah, "Topic modelling and hotel rating prediction based on customer review in Indonesia," International Journal of Management and Decision Making, vol. 20, no. 3, pp. 282–307, 2021.

[16] F. Hawlitschek, T. Teubner, and C. Weinhardt, "Trust in the sharing economy," Die Unternehmung, vol. 70, no. 1, pp. 26–44, 2016. doi: 10.5771/0042-059X-2016-1-26.

[17] E. Ert, A. Fleischer, and N. Magen, "Trust and reputation in the sharing economy: The role of personal photos in Airbnb," Tour. Manag., vol. 55, pp. 62–73, 2016, doi: 10.2139/ssrn.2624181.

[18] D. Demirol, R. Das, and D. Hanbay, 'A key review on security and privacy of big data: issues, challenges, and future research directions', SIViP, Sept. 2022, doi: 10.1007/s11760-022-02341-w.

[19] D. Demirol, R. Das, and D. Hanbay, 'Büyük veri üzerine perspektif bir bakış', in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Sept. 2019, pp. 1–9. doi: 10.1109/IDAP.2019.8875902.

[20] Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. In Proceedings of the 3rd International Balkan Conference on Communications and Networking (BalkanCom). Istanbul, Turkey.

[21] R. Güran, "Stopword list for Turkish," GitHub, [Online]. Available: https://github.com/stopwords-iso/stopwords-tr.

[22] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in Int. Conf. Learn. Represent. (ICLR), 2020. doi: 10.48550/arXiv.2003.10555.

[23] M. Laurer, *mDeBERTa-v3-base-xnli-multilingual-nli-2mil7* [Model]. Hugging Face, 2021. [Online]. Available: https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7.

[24] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008. doi: 10.1017/CBO9780511809071.

[25] F. Nurifan, R. Sarno, and K. R. Sungkono, "Aspect based sentiment analysis for restaurant reviews using hybrid elmo-wikipedia and hybrid expanded opinion lexicon-senticircle," International Journal of Intelligent Engineering and Systems, vol. 12, no. 6, pp. 47–58, 2019, doi: 10.22266/ijies2019.1231.05.

[26] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed. New York, NY, USA: Springer, 2021. doi: 10.1007/978-1-0716-1418-1.