BMBAD BJESR Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches

Large Group Decision Making for Aspect-Level Consensus Evaluation in Low-Rated App Reviews

Düşük Puanlı Uygulama Yorumlarında Özellik Bazında Görüş Birliği Değerlendirmesi için Büyük Grupla Karar Alma

[1]Ahmet Cumhur ÖZTÜRK

[1]Aydın Adnan Menderes University, Söke Vocatinal School, Department of Computer Technologies, Söke, Aydın, Türkiye

[1]cumhur.ozturk@adu.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received :11 June 2025
Accepted : 29 September 2025

Keywords:

Large Group Consensus, Aspect Based Sentiment Analysis, Sentence Embedding, Cosine Similarity, Online Reviews

ABSTRACT

Consumer to consumer (C2C) e-commerce platforms allow users to buy and sell second hand products and they offer affordability and support sustainable consumption. In these environments, user generated reviews provide valuable insights into service failures. Traditional sentiment analysis and Aspect Based Sentiment Analysis (ABSA) methods primarily focus on classifying the polarity of opinions expressed in reviews. However, these approaches often fall short in capturing the user agreement or identifying whether specific complaints are widely shared.

The present study adopts a Large Group Decision Making framework to analyze low rated Turkish language reviews from a second hand marketplace app. The approach integrates ABSA and semantic similarity modeling to improve the interpretability of user complaints. Also it enables to detect widely shared and divergent complaints and also offers more actionable insights than traditional sentiment aggregation.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 11 Haziran 2025
Kabul : 29 Eylül 2025

Anahtar Kelimeler:

Büyük Grup Karar Verme, Hedef Tabanlı Duygu Analizi, Cümle Vektörü, Kosinüs Benzerliği, Online Yorumlar

ÖZET

Tüketiciden tüketiciye (C2C) e-ticaret platformları, kullanıcıların ikinci el ürünleri alıp satmasına olanak tanımaktadır ve bu sayede kullanıcılara uygun fiyatlı alışveriş imkânı suna ve sürdürülebilir tüketime katkı sağlamaktadır. Kullanıcılar tarafından oluşturulan yorumlar ile bu platformlarda ortaya çıkan hizmet aksaklıklarını ortaya çıkarmak mümkündür. Geleneksel duygu analizi ve Hedefe Dayalı Duygu Analizi (ABSA) yöntemleri, genellikle yorumlardaki ifadelerin olumlu, olumsuz ya da nötr şeklindeki gruplandırılmasına odaklanmaktadır. Ancak bu yaklaşımlar, kullanıcılar arasındaki fikir birliğini yakalamakta ya da belirli şikâyetlerin ne kadar yaygın olduğunu tespit etmede yetersiz kalabilmektedir.

Bu çalışma, ikinci el bir e-ticaret uygulamasından alınan düşük puanlı Türkçe yorumları analiz etmek için Büyük Grup Karar Verme (Large Group Decision Making - LGDM) uygulamaktadır. Bu yaklaşım, ABSA ve anlamsal benzerlik modellemesini entegre ederek kullanıcı şikâyetlerinin yorumlanabilirliğini artırmayı amaçlamaktadır. Ayrıca, yaygın şekilde paylaşılan ya da ayrışan şikâyetleri tespit edebilmekte ve geleneksel duygu toplulaştırma yöntemlerine kıyasla daha uygulanabilir içgörüler sunmaktadır.

ORCID ID: [1]0000-0002-2677-3269

1. INTRODUCTION

The rise of mobile application based consumer to consumer (C2C) e-commerce platforms has dramatically expanded engagement in second hand product exchange. These platforms offer users convenience, cost savings and opportunities to participate in circular consumption. However, they also present challenges that differ from traditional e-commerce platforms, such as issues related to transaction safety, operational reliability and platform trust [1]. In these C2C environments, where sellers are unverified and product descriptions are user-generated, dissatisfaction frequently stems from failures such as fraudulent listings, delivery problems, and unresponsive support systems. User generated online reviews serve as a rich source of information for understanding platform shortcomings. Prior research has shown that such reviews influence purchasing decisions [2] and contain detailed insights into specific pain points experienced by users [3]. However, much of the existing work on online reviews in e-commerce either focuses on general sentiment or fails to distinguish between multiple issues mentioned within a single review [4]. This limitation is especially critical in second-hand marketplaces, where reviews often contain a mix of praise and complaints about different features or interactions. Moreover, it has been widely documented that consumers are more influenced by negative reviews than positive ones and that negative feedback offers more actionable information for improving service and platform design [5,6]. Despite this, relatively few studies have focused specifically on negative reviews, particularly in the context of second hand mobile commerce. Even fewer have applied advanced text mining techniques to isolate the specific elements of a platform that elicit dissatisfaction [2].

Large Group Decision Making (LGDM) is a decision-making framework designed to aggregate and analyze the opinions of a large and diverse population. Traditional review analysis methods often rely on document level sentiment aggregation which risks overlooking the complexity and inconsistency of user feedback. This study adopts an LGDM perspective to evaluate and aggregate user complaints from low rating Turkish language reviews collected from Dolap, a second-hand marketplace application on Google Play. To extract fine grained, aspect specific concerns, reviews were first segmented into individual sentences. These sentence level expressions were then grouped into clusters of shared concerns, allowing the identification of common dissatisfaction themes and the degree of consensus within each aspect. By integrating clustering, LGDM and semantic similarity modeling, this study offers a scalable and interpretable framework for analyzing user complaints in C2C platforms. The findings provide actionable insights for platform managers by highlighting the most critical service failures consistently expressed across users, as well as those requiring more nuanced investigation. To systematically explore the nature of user dissatisfaction and guide the analytical process, the following research questions were formulated:

RQ1: How can aspects of low rating feedback automatically be extracted from Turkish language reviews on second hand marketplace applications?

RQ2: How can aspect specific dissatisfaction expressed by the majority of users be identified through semantic consensus modeling?

To answer Question 1, the Turkish reviews were segmented into individual sentences and embedded using SBERT [7]. These embeddings were then clustered using unsupervised techniques and each cluster was labeled using the top TF-IDF keywords it contained. This enabled the automatic extraction of distinct aspects of negative feedback without requiring manual annotation. To answer Question 2, sentence embeddings were compared with aspect-level consensus vectors using cosine similarity. This allowed the study to quantify how strongly user complaints converged around shared themes and revealed dominant types of dissatisfaction in each aspect.

2. LITERATURE REVIEW

Large group decision making (LGDM) refers to exploring a collective agreement among numerous users through structured methods for identifying shared priorities or solutions [8]. It has emerged as a powerful framework for aggregating and analyzing user opinions when traditional decision-making assumptions, such as initial consensus or expert input, are not available. In the literature, many studies have conducted LSGDM on online reviews for achieving various tasks such as service evaluation, trust assessment and failure detection [3, 9]. Ji et al. [9] developed a minimal variance weight based LGDM model to fairly allocate weights to peer-to-peer accommodation users and extracted key experiential themes from online feedback. Yuan et al. [3] proposed a dual fine-tuning LSGDM model that integrates sentiment analysis with density based clustering algorithms to improve consensus convergence in large scale user feedback. Wu et al. [10] proposed an LGDM method based on aggregated user sentiments from online reviews to assist users in making purchase decisions. Ma et al. [11] used SBERT embeddings and cosine similarity to evaluate overall cruise satisfaction by clustering reviews based on semantic similarity, focusing on identifying key service attributes rather than analyzing sentence-level consensus within specific complaint categories. Shi et al. [12] applied aggregated sentiment scores to model large group decision making in cruise reviews, using cosine similarity solely for dictionary expansion to enrich sentiment lexicons, without leveraging it for semantic similarity comparisons between user review sentences.

In natural language processing (NLP), sentence embeddings are dense vector representations that capture the semantic meaning of entire sentences. Unlike word embeddings, sentence embeddings aim to encode syntax, semantics, and context into a single vector representation. Traditional sentence embedding approaches, such as

Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), represent sentences as high-dimensional sparse vectors based on word occurrences and frequencies [13]. Although these approaches are simple to implement and computationally efficient, they lack the ability to capture word order and syntactic structure. Neural network based models, such as SkipThought [14] and InferSent [15], use recurrent neural networks to learn distributed sentence representations. These models outperform traditional methods by enabling more context-aware sentence representations through capturing word order and syntactic structures. More recently, Transformer-based models like BERT [16], Sentence-BERT [7], and SimCSE [17] have achieved state-of-the-art performance in producing context rich and semantically meaningful sentence embeddings.

Cosine similarity is a metric used to measure the cosine of the angle between two non-zero vectors in an inner product space. It effectively captures vector orientation, making it particularly suitable for comparing sentence embeddings and assessing semantic similarity regardless of sentence length. In the context of e-commerce, combining sentence embeddings with cosine similarity has proven effective in various applications, such as product matching and sentiment analysis. For instance, a method for clustering aspect phrases in e-commerce reviews to facilitate the identification of common customer concerns without manually labeling textual data was performed in [18]. Similarly, Saha [19] evaluated the impact of various text embeddings on clustering performance and highlighted the effectiveness of sentence embeddings in this regard. Zhou et al. [20] proposed a decentralized multipartite consensus model using SBERT and cosine similarity to enhance large group decision-making. UI Haq and Fraca [21] applied sentence embeddings with cosine similarity to evaluate idea novelty and consensus in collaborative problem-solving contexts.

While second hand marketplace platforms have been increasingly examined for trust dynamics [22], sustainability, and peer-to-peer interactions [23], their integration with structured decision-making frameworks like Large Group Decision Making (LGDM) remains underexplored. Additionally, although recent NLP advancements have demonstrated the effectiveness of SBERT embeddings in clustering long texts and modeling semantic similarity [24], these techniques have not yet been applied to model sentence level consensus in large scale, non-English consumer-to-consumer (C2C) review environments, such as Turkish.

This study introduces a novel framework that combines SBERT based sentence embeddings, aspect level clustering, semantic similarity filtering and consensus scoring. While prior research has explored individual components of this methodology such as SBERT based clustering [19] or consensus modeling in structured LGDM contexts [20, 21], none of them have integrated all three elements (clustering, similarity filtering and consensus scoring) into a unified framework tailored for aspect based complaint aggregation. Also existing approaches typically focus on English language datasets or domain specific tasks such as cruise satisfaction [11] or dictionary expansion [12]. In contrast, this study introduces an unsupervised LGDM based framework that clusters user complaints, filters them based on semantic alignment and quantifies consensus using SBERT and cosine similarity all within the context of Turkish language, low-rated second hand platform reviews. Key innovations include the explicit quantification of the trade-off between semantic alignment and the proportion of sentences retained within each aspect cluster. No prior study appears to apply such a combination to large-scale, unstructured consumer review analysis in a non-English mobile app setting.

3. METHODOLOGY

This study aims to identify and aggregate the most prominent user complaints in second hand marketplace applications by applying a sentence level, large group decision making framework. The workflow of the study is shown n Figure 1. First, reviews and their corresponding star ratings are collected from GooglePlay. Then in the Data Preprocessing stage, user reviews with low ratings (1–2 stars) were filtered and segmented into individual sentences with using SpaCy multilingual model (xx_ent_wiki_sm) to capture fine grained expressions of dissatisfaction. Next Zemberek [25] based lemmatization and regular expression filtering was applied. The cleaned sentences are then converted into vector representations using a pretrained SBERT model. In the Aspect Label Assignment phase aspect related sentence clusters are generated through a two-step process clustering process; KMeans is applied to form fine-grained clusters and then these clustered are merged using agglomerative clustering. Each resulting cluster is interpreted by extracting its most representative terms using TF-IDF for aspect label assignment. In the Consensus Degree Computation phase to evaluate the degree of semantic agreement within each aspect a consensus computation is average cosine similarity between each sentence embedding and the mean sentence embedding of its cluster. This structured approach allows for a scalable and interpretable analysis of negative user experiences in the second-hand market domain.

3.1 Data Collection and Preprocessing

Online customer reviews written in Turkish language posted between January 1, 2022, and December 31, 2024, were collected from the Google Play Store for the Dolap mobile application, a widely used second-hand marketplace platform in Turkey. For each review, the review text, posting date and star rating of the review were collected using Selenium package of Python programming language. Table 1 presents a sample from the collected review dataset, the dataset contains columns for review ID, review content, user rating and the date of submission. Consistent with prior research that uses review rating stars to infer sentiment polarity [26], reviews with 1–2-star

ratings were assumed to reflect negative user experiences while those with a rating of 3 were considered neutral and those with a rating of 4 or higher were considered positive. A total of 17,628 online reviews were collected and 9,377 were classified as negative where 8,422 had a rating of 1 and 955 had a rating of 2.

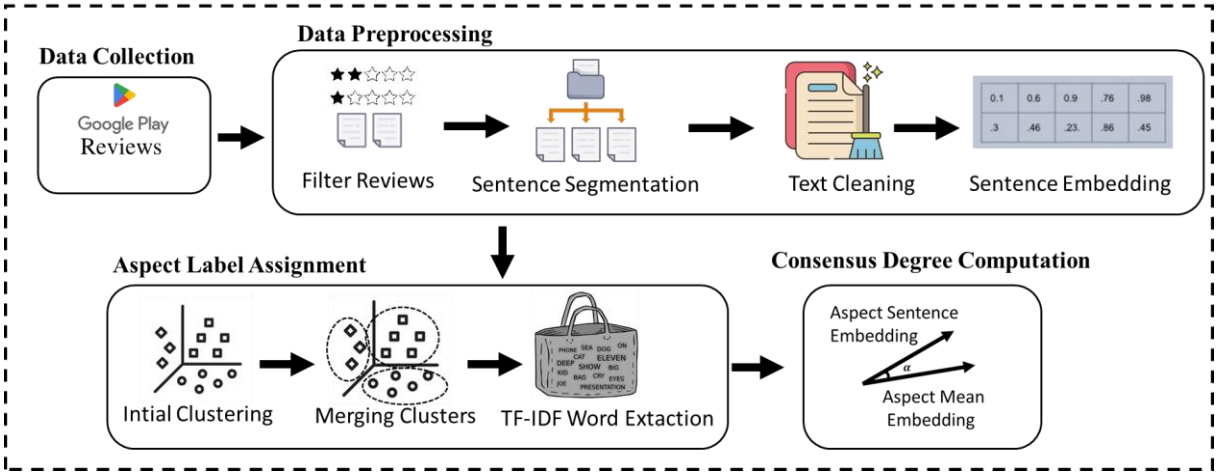


Figure 1. Workflow of the large group decision making (LGDM) framework for complaint consensus analysis.

Table 1. A sample of review dataset.

ID	Review	Rating	Date
1	(TR[a]) Aldığımız ürünü iade ettiğimiz zaman kendi adığı komisyonu geri iade etmeyen uygulama resmen gaspçılık yapıyorlar (EN[b]) The practice of not returning the commission it received when we return the product we purchased is literally extortion.	1	01.10.2023
2	(TR[a]) Uygulamaya giriş yapamıyorum sürekli hata veriyor . (EN[b]) I can't log in to the app, it keeps giving me an error	2	01.16.2023
3	(TR[a]) Uygulamanın arsivledigi ürünler için bir bölüm yapilirsa çok iyi olur dönüş olursa teşekkürler (EN[b]) It would be great if there was a section for the products I archived in the application. Thanks for your feedback.	3	01.17.2023
4	(TR[a]) Harika bir uygulama son derece güvenilir fakat hizmet bedeli kesiyor biraz indirip bir bakın derim (EN[b]) It's a great app, extremely reliable, but it charges a service fee. I'd say download it and take a look.	4	02.02.2023
5	(TR[a])Tşk ederim güzel güvenilir bir site 👍 (EN[b]) Thank you, it is a nice and reliable site 👍	5	01.09.2023

[a]Turkish,[b]English

The quality and relevance of the textual data were ensured through a multi-step preprocessing pipeline tailored for Turkish-language user reviews. First, sentence segmentation was performed using the SpaCy natural language processing library, which tokenizes Turkish language reviews into individual sentences based on language specific syntactic rules**.** Language detection was then performed using the Google Translate API at the sentence level and only sentences identified as Turkish were retained. Among these, those containing two or more foreign (non-Turkish) words were further excluded. For instance, the sentence *"Email adresim silindi lütfen yardım edin"* contains only one foreign word ("email") and was retained, whereas *"Login error veriyor password kabul etmiyor"*, which includes multiple foreign words, was removed. Next, sentences with fewer than five words were discarded to remove noisy or uninformative content. Remaining sentences were cleaned by removing emojis, numbers, URLs and extraneous punctuation. Repetitive characters (e.g., "çooook") were normalized by reducing any character repeated more than twice to a single instance (e.g. "çok") with using a regular expression specifically designed for repetition handling and then all text was converted to lowercase. After this, the text underwent a custom Turkish normalization process that included Unicode normalization, regular expression based noise removal, stopword filtering and lemmatization using Zemberek [25]. While Zemberek effectively reduced most morphological variation, manual inspection revealed some recurring word forms that required additional normalization. Therefore, a small predefined normalization dictionary was created to map frequently occurring variants to their common root forms. For example, words such as "hesabım" (my account), "hesabınıza" (to your account), and "hesaplar" (accounts) were all mapped to the root form "hesap," ensuring semantic consistency

across reviews. Words shorter than three characters were also removed and UTF-8 encoding was enforced throughout the pipeline.
This comprehensive preprocessing pipeline ensured that the remaining sentences were linguistically meaningful, normalized and morphologically simplified for enhancing the quality of sentence embeddings and improving the performance of clustering and similarity based. The SBERT was used to generate semantically meaningful sentence embeddings, and it has a maximum input capacity of 512 tokens. Sentence embeddings were generated using the paraphrase-multilingual-MiniLM-L12-v2 model, a lightweight multilingual variant of SBERT optimized for semantic similarity tasks. This model was preferred due to its proven effectiveness in multilingual embedding benchmarks, its support for a wider variety of language constructs, and its better compatibility with sentence level representation tasks. Since SBERT has a maximum input length of 512 tokens, the token length of each review sentence was calculated in advance to ensure that no truncation would occur during embedding generation. To ensure compatibility with the input constraints of SBERT model, the token length of each review sentence was calculated before sentence embedding. Figure 2 shows the distribution of token lengths in preprocessed sentences where the x-axis represents the number of tokens per sentence and the y-axis represents their frequency. As illustrated in Figure 2 none of the review sentence length exceeded 512 token limit and this ensures that the full content of each review is preserved and analyzed.

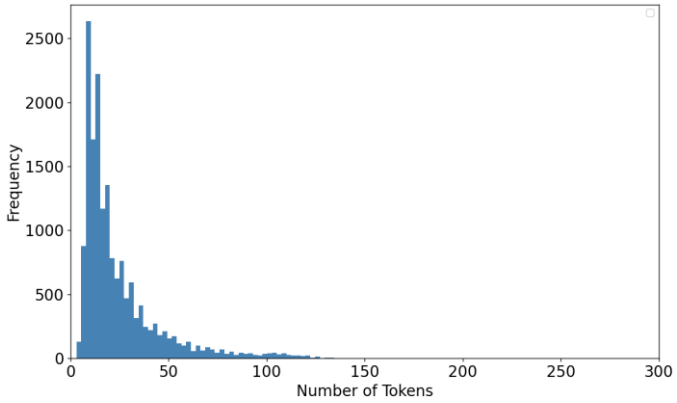


Figure 2. Distribution of tokens in each negative review sentence.

3.2 Semantic Clustering of Complaints for Aspect Identification

The aspects of reviews are essential dimensions for understanding customer dissatisfaction in large group decision making (LGDM) context. Aspects were automatically extracted from the negative review corpus with an unsupervised aspect extraction pipeline as follows; Following the preprocessing phase, the sentence embeddings were generated using the pre trained "paraphrase-multilingual-MiniLM-L12-v2" model from the SentenceTransformers library [6] that provides multilingual sentence level. These high dimensional semantic representations were then clustered with using a two-step procedure. First, K-Means clustering with k=50 was used to partition the embeddings into fine grained groups. Subsequently, the centroid vectors of these clusters were subjected to agglomerative hierarchical clustering (average linkage, cosine metric) to merge semantically similar clusters and as a result the number of cluster size was reduced to seven. To enhance the interpretability, a label was assigned to each cluster based on the top ranked terms that were identified using term frequency inverse document frequency (TF-IDF) scores. In Figure 3 the semantic structure of the clustered sentence embedding are shown with a two dimentional t-SNE visualization of the seven clusters. In this figure each point corresponds to a sentence and each color represents one of the semantically coherent clusters. The aspect labels assigned to these clusters are as follows: Platform Comparison / Usage Preference, Low Value for Money, Customer Support Failures, Negative Recommendation, Shipping & Fee Complaints, Login & Access Issues and Low Ratings & Fairness Concerns.
The cluster labels, top keywords in TF-IDF and description of each label is shown in Table 2. In this table the Cluster Label column presents the manually assigned names for each cluster based on the interpretation of high-ranking TF-IDF terms. The Top Keywords (TF-IDF) column lists the most representative terms extracted using Term Frequency Inverse Document Frequency (TF-IDF) analysis and the Description column provides an explanation of the rationale behind each cluster label.

3.3 Consensus Measurement within Aspect Groups

In the context of negative online reviews, consensus refers to the degree of agreement among users in how they express complaints about a specific topic within a given aspect. While all reviews grouped under the same aspect are negative, sub issues or topics mentioned within reviews can vary. For example, within the customer service aspect while some users may criticize slow responses, others may criticize about unresolved issues. Consensus

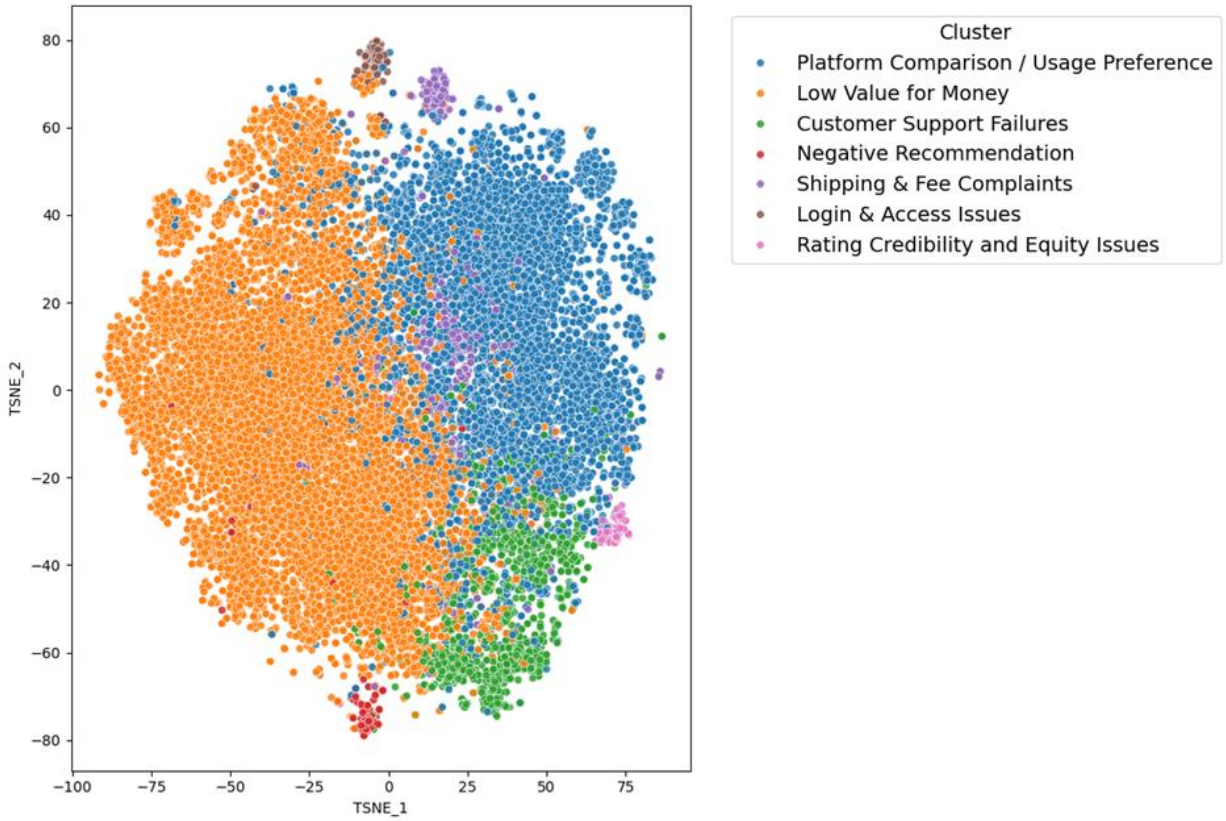


Figure 3. t-SNE visualization of auto merged clusters and their names.

Table 2. Identified complaint clusters with TF-IDF keywords.

Cluster Labels	Top Keywords (TF-IDF)	Description
Low Value for Money	(TR[a]) kargo, komisyon, fazla, bedel, hizmet (EN[b]) shipping, commission, excessive, cost, service	Users expressing dissatisfaction with the product or service relative to its cost.
Shipping & Fee Complaints	(TR[a]) gönderim, fiyat, kargo, uygulama, teslimat (EN[b]) shipping, price, shipping, app, delivery	Complaints about delivery delays, unexpected shipping fees or courier issues.
Login & Access Issues	(TR[a]) giriş, hesap, şifre, yapmak, doğrulamak (EN[b]) login, account, password, perform, verify	Complaints related to problems signing in, password recovery, or app access.
Customer Support Failures	(TR[a]) destek, hizmet, ulaşmak, telefon, temsilci (EN[b]) support, service, reach, phone, representative	Issues concerning the quality or availability of support from platform staff.
Platform Comparison / Usage Preference	(TR[a]) platform, gardrops, alternatif, memnun, iade (EN[b]) platform, gardrops, alternative, satisfied, refund	Mentions of users switching to or recommending other platforms.
Negative Recommendation	(TR[a])tavsiye, önermek, asla, memnuniyetsizlik, pişmanlık (EN[b]) recommend, suggest, never, dissatisfaction, regret	Strongly negative opinions discouraging others from using the platform.
Low Ratings & Fairness Concerns	(TR[a]) yanıltıcı, hizmet, adil, temsilci, değerlendirme (EN[b]) misleading,customer, service, fair, representative, rating	Complaints about perceived unfairness or manipulation in the platform's review or rating system

[a]Turkish,[b]English

measurement evaluates whether users are converging on the same specific complaint or expressing a wide range of grievances under that aspect. High consensus indicates strong alignment in user dissatisfaction and low consensus reflects more scattered or individualized concerns.

To evaluate consensus within each aspect, the sentence embeddings previously generated using SBERT during the clustering phase were reused. The sentence embeddings within each cluster were averaged separately to compute a consensus vector representing the central semantic tendency of each aspect. Then, cosine similarity between each individual sentence embeddings and consensus vector within each aspect group was calculated. The resulting similarity score indicates how closely a sentence aligns with the dominant theme of its cluster where values closer to 1 denotes high semantic alignment and values closer to 0 reflects divergence. For quantifying the overall agreement within each aspect group, a consensus degree metric was used.

Algorithm 1. Large group decision making algorithm.

Input: A set of review sentences S, where each sentence s ∈ S includes:
- Sentence embeddings (SBERT) : s.embedding
- Assigned aspect label: s.aspect

Output: The consensus score and size for a specific aspect cluster aspectConsensusScores.

1 A= {a_1, a_2, ..., a_n} from S ////Identify all unique aspects
2 for each aspect a ∈ A:
3 Let $V_a = \{s.\text{embedding} \mid s \in S \wedge s.\text{aspect} = a\}$ //// Embedding vectors for aspect a
4 if $|V_a|$>0:
5 Let $\overline{v_a} = (1/|V_a|)\sum_{i=1}^{|V_a|} v_i$
6 agree_count=0
7 for i=1 to $|V_a|$:
8 $sim_i = cosine(v_i, \overline{v_a})$ ////Compute similarity
9 if $sim_i \geq 0.7$:
10 agree_count=agree_count+1
11 C_a = agree_count/$|V_a|$ ////compute consensus degree
12 Add to aspectConsensusScores: (a, $|V_a|$, C_a) //// aspect label, sentence count, consensus score

This metric is defined as the proportion of sentences within an aspect cluster that have a cosine similarity score of 0.70 or higher with the cluster's consensus vector. The threshold of 0.70 was selected based on previous studies in semantic similarity tasks where values above 0.70 typically indicate strong conceptual alignment between sentences [20]. Algorithm 1 shows the algorithm for computing aspect level consensus degree within each aspect group. The algorithm takes as input the sentence embeddings and their corresponding aspect labels and outputs the consensus scores for each aspect group. In this algorithm, first all unique aspects in the given dataset were identified (Step 1). Next, sentence embeddings corresponding to each aspect were grouped together (Step 2). Through Steps 3 to 5, the average of these embeddings was computed for each aspect to obtain the consensus vector representing the central semantic tendency. Steps 6 through 10 counted the number of sentence embeddings within each aspect group whose cosine similarity score with the consensus vector was greater than or equal to 0.70. In Steps 11, the proportion of aligned sentences was calculated and stored along with the aspect label and sentence count to the aspectConsensusScores variable.

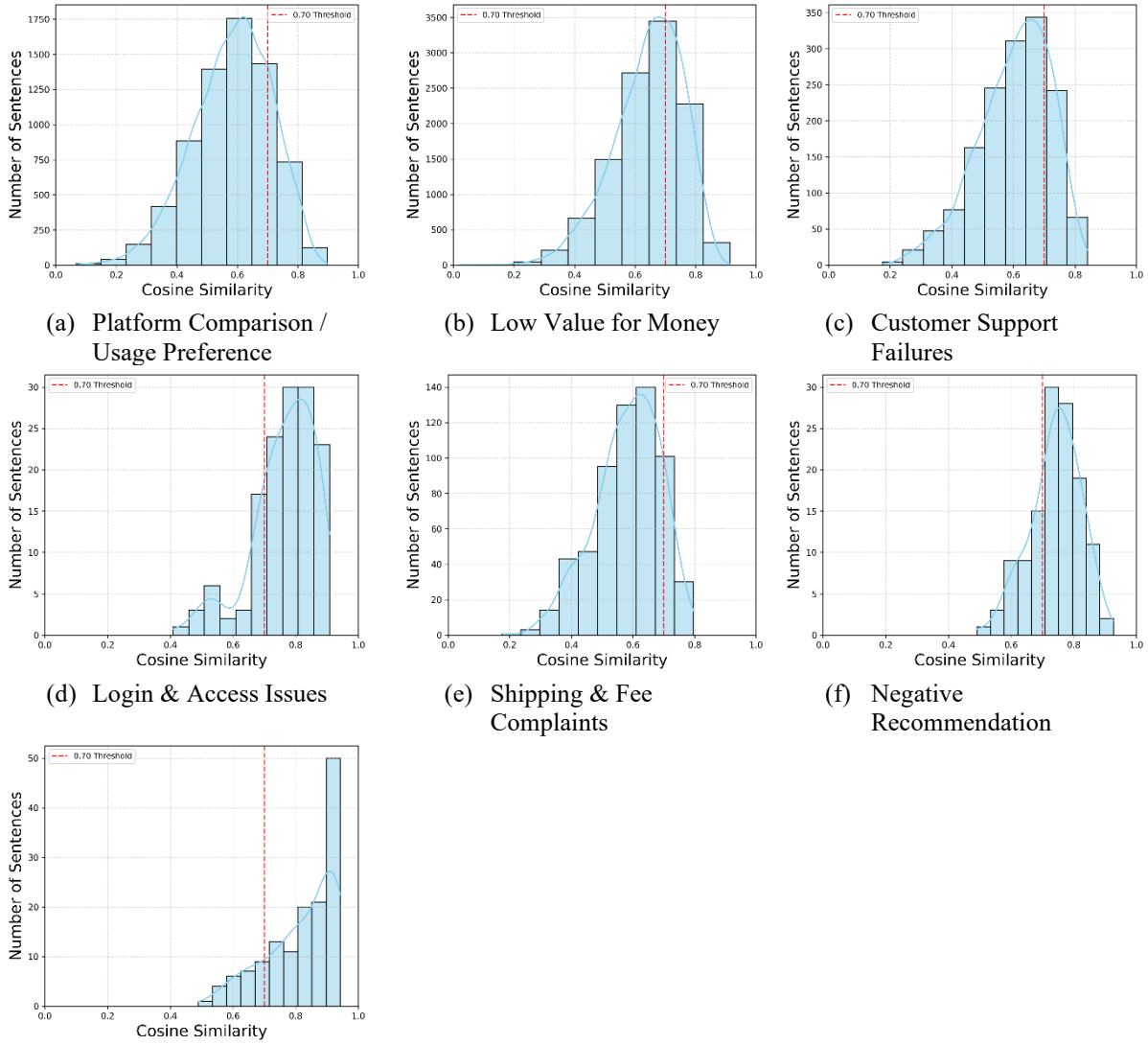
Traditional ABSA tasks often aim to identify what users talk about and how they feel by applying polarity classification or measuring aspect term frequency. However, such methods typically fail to capture the degree of semantic alignment or shared concern among users within an aspect. In contrast, this study measures the semantic similarity between review sentences and their aspect centroid to detect coherent complaint patterns. The originality of the approach lies in applying a cosine similarity threshold ($\geq$ 0.70) to filter off-topic sentences and computing a consensus score to quantify user alignment within each aspect cluster. While traditional LGDM approaches rely on structured inputs such as expert evaluations, scoring matrices or preference aggregation models [27], this study extends the LGDM framework to unstructured user-generated text by leveraging SBERT-based sentence embeddings and cosine similarity. This enables semantic-level agreement detection in open-ended reviews, offering a replicable, unsupervised method for identifying widely shared concerns in Turkish low-rated app reviews and addressing limitations of both polarity-based ABSA and conventional LGDM.

4. RESULTS

4.1 Consensus Threshold Selection and Sensitivity Analysis

The degree of semantic consensus among user-generated sentences in large group decision making (LGDM) depends on selecting an appropriate similarity threshold. Although previous studies commonly adopt a threshold of 0.70, the optimal value may vary depending on the structure and diversity of the review dataset. To validate the suitability of the 0.70 threshold in the context of this study, a threshold sensitivity analysis was conducted and the results are presented in Table 3. This table displays the top five TF-IDF keywords extracted from review sentences across similarity thresholds ranging from 0.60 to 0.80. As the threshold increases, the extracted keywords become more semantically homogenous, indicating stronger alignment of user concerns. However, this increase in homogeneity comes at the cost of reduced sentence coverage, as shown in Figure 4. In each subplot of Figure 4, the red dashed line indicates the 0.70 similarity threshold applied for consensus filtering. While the majority of aspects retain a reasonable number of sentences at higher thresholds, aspects such as Customer Support Failures and Login & Access Issues show a steep decline beyond the 0.80 mark. In contrast, aspects like Shipping & Fee

Complaints still preserve sufficient sentences for analysis even at the 0.80 level, allowing extraction of meaningful keywords such as “kargo” (shipping), “çok” (very), and “ücret” (fee).



(a) Platform Comparison / Usage Preference

(b) Low Value for Money

(c) Customer Support Failures

(d) Login & Access Issues

(e) Shipping & Fee Complaints

(f) Negative Recommendation

(g) Low Ratings & Fairness Concerns

Figure 4. Distribution of sentence level cosine similarities within each complaint aspect: (a) Platform Comparison / Usage Preference, (b) Low Value for Money, (c) Customer Support Failures, (d) Login & Access Issues, (e) Shipping & Fee Complaints, (f) Negative Recommendation, and (g) Low Ratings & Fairness Concerns.

The strength of the 0.70 threshold is supported by both the representativeness of the extracted keywords in Table 3 and the sentence similarity distributions in Figure 4. For example, in the Low Value for Money aspect, the keywords at the 0.70 threshold include “pahalı” (expensive), “komisyon” (commission), “değer” (worth), “fiyat” (price), and “uygun değil” (not fair), reflecting coherent user concerns about excessive pricing and perceived unfair value. At the stricter 0.80 threshold, the keywords shift toward more emotionally charged and severe terms such as “kazık” (rip-off), “soygun” (robbery), “değmez” (not worth), “komisyon” (commission), and “fahiş” (inflated), indicating a narrower and more intensely dissatisfied subset of reviews.

The richness of user perspectives diminishes as review sentence coverage decreases, which in turn restricts the analytical depth of the extracted keywords. At the 0.70 threshold, there appears to be a balanced trade-off between meaningful semantic alignment and adequate coverage. This balance is visually supported by Figure 4, where most sentence similarities fall between the 0.60 and 0.80 thresholds, with few exceeding 0.80. Therefore, the 0.70 threshold is deemed optimal for this dataset, effectively balancing broad inclusion with focused consensus.

4.2 Consensus Analysis

As explained in Section 3.3, consensus within each aspect was calculated based on the proportion of sentences within each aspect group whose cosine similarity to the consensus vector exceeded the 0.70 threshold. The results of the analysis are presented in Figure 5.

Table 3. Top 5 TF-IDF keywords extracted from review sentences at various similarity thresholds.

	Aspect Threshold		
Aspect Name	**0.60**	**0.70**	**0.80**
Platform Comparison / Usage Preference	(TR[a])gardrops, var, kullanın, uygulamalara, için (EN[b]) gardrops, exists, use, apps, for	(TR[a])gardrops, başka, uygulamalara, öner, geçiyorum (EN[b]) gardrops, other, apps, suggest, I switch	(TR[a]) sil, iyi, haram, soygun, zıkkım (EN[b]) delete, good, illegitimate, robbery, cursed
Low Value for Money	(TR[a]) fiyat, ürün, pahalı, ücret, gönderim (EN[b]) price, product, expensive, fee, shipping	(TR[a]) pahalı, komisyon, değer, fiyat, uygun değil (EN[b]) expensive, commission, worth, price, not fair	(TR[a]) kazık, soygun, değmez, komisyon, fahiş (EN[b]) rip-off, robbery, not worth, commission, inflated
Customer Support Failures	(TR[a]) destek, ulaşmak, temsilci, hizmet, cevap (EN[b]) support, reach, representative, service, respond	(TR[a]) destek, ulaşmak, aramak, açmak, bekletmek (EN[b]) support, reach, call, answer, wait	(TR[a]) umursamak, ilgilenmek, haketmek, cevapsız, mağdur (EN[b]) ignore, care, deserve, unresponsive, victim
Login & Access Issues	(TR[a])hesap, kapatmak, giriş, telefon, yapmak (EN[b]) account, close, login, phone, do	(TR[a])hesap, kapatmak, silmek, giriş, yer (EN[b]) ccount, close, delete, login, place	(TR[a])kapatmak, açmak, tekrar, giriş, silmek (EN[b]) close, open, again, login, delete
Shipping & Fee Complaints	(TR[a]) kargo, fiyat, uygulama, ürün, gönderim (EN[b]) shipping, price, app, product, delivery	(TR[a]) kargo, ücret, beklemek, geç, sorun (EN[b]) shipping, fee, wait, late, issue	(TR[a]) komisyon, kazık, soygun, değmez, fahiş (EN[b]) commission, rip-off, robbery, not worth, overpriced
Negative Recommendation	(TR[a]) tavsiye, memnuniyetsizlik, pişmanlık, yorum, şikayet (EN[b]) recommend, dissatisfaction, regret, review, complaint	(TR[a]) asla, önermek, şikayet, pişmanlık, tekrar (EN[b]) never, suggest, complaint, regret, again	(TR[a]) tavsiye, önermek, pişmanlık, değmez, asla (EN[b])recommend, suggest, regret, not worth, never
Low Ratings & Fairness Concerns	(TR[a])değerlendirme, hizmet, müsteri, puan, yorum (EN[b]) rating, service, customer, score, review	(TR[a]) adil, temsilci, yanıltıcı, düşük, yıldız (EN[b]) fair, representative, misleading, low, star	(TR[a]) adil, haksız, yanlı, güvenilmez, sahte (EN[b]) fair, unfair, unreliable, fake,

[a]Turkish,[b]English

In this figure the x-axis represents the proportion of sentences within each aspect whose cosine similarity to the consensus vector exceeded the 0.70 threshold and the y-axis shows the aspect names along with the number of sentences they contain. As shown in the figure, aspects "Low Ratings & Fairness Concerns" (0.82), "Login & Access Issues" (0.78) and "Negative Recommendation" (0.72) exhibit the highest consensus degrees. This indicates that users expressing complaints in these categories tend to articulate similar concerns, reflecting well defined and commonly shared dissatisfaction. These issues are expressed clearly and consistently by users this makes them high priority problems that are generally easier to address and manage. In contrast, aspects like "Shipping & Fee Complaints" (0.13) and "Platform Comparison / Usage Preference" (0.19) show much lower consensus, suggesting greater variation or ambiguity in how users describe their issues. These weak consensus groups may require more nuanced analysis or refinement into sub aspects before actionable insights can be derived. Within LGDM, such low consensus issues may benefit from additional user segmentation, attribute weighting or follow-up studies to disentangle the varied concerns they contain.

4.3 Comparative Analysis of Semantic Alignment in Aspect Clusters

To evaluate the effectiveness of the proposed LGDM framework, we compared it against an embedding based aspect clustering baseline that uses SBERT sentence embeddings for unsupervised review grouping [12]. This baseline method performs aspect level clustering without applying any cosine similarity threshold for sentence selection. It retains all review sentences in a given aspect cluster, regardless of how semantically close they are to the core theme. In contrast, our method introduces a cosine similarity threshold (≥ 0.70) to retain only sentences that are semantically aligned with their corresponding aspect centroid known as a vector representation computed as the average of all sentence embeddings within an aspect. For both methods aspect centroid of all sentence embeddings belonging to an aspect is calculated and then each sentence s in an aspect cluster its cosine similarity to the centroid vector c is calculated using the following formula [13];

$$CosSim(s_i, c) = \frac{\vec{s}\ \vec{c}}{\|\vec{s}\|\ ||\vec{c}||} \quad (1)$$

where $\vec{s}$ represents the sentence embedding vector, $\vec{c}$ denotes the aspect centroid vector and $\|\vec{s}\|\ ||\vec{c}||$ is the product

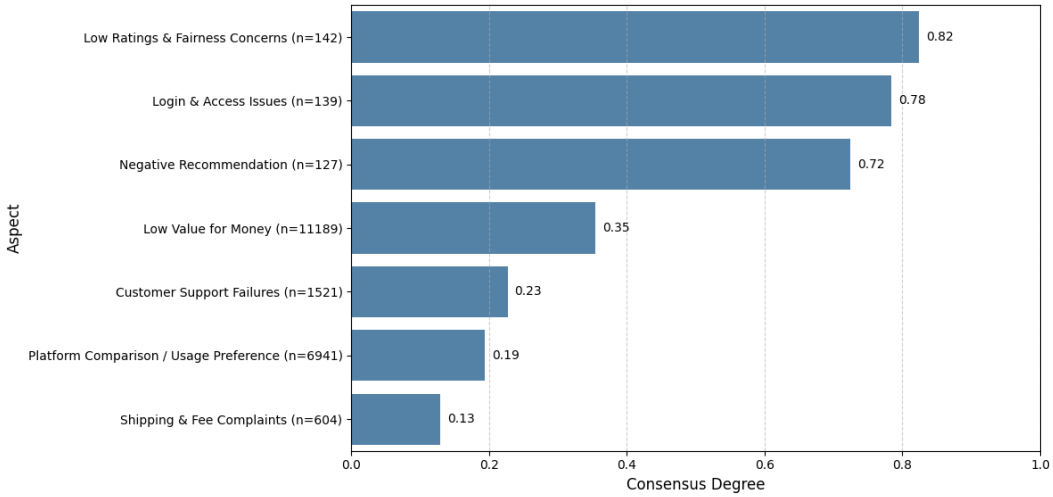


Figure 5. Cosine similarity-based consensus degree within each aspect.

of their Euclidian norms. The resulting similarity score ranges from -1 to +1 where values closer to +1 indicate stronger semantic similarity.

To assess the internal thematic alignment within each aspect cluster, the average cosine similarity between all sentences and their aspect centroid was calculated in the baseline method. For the LGDM approach, the consensus score was computed as the average cosine similarity of only those sentences that exceeded the predefined threshold (≥ 0.70).

Table 4. Semantic alignment and consensus metrics across complaint aspects.

Aspect	Average Cos.Sim.	Consensus	Consensus Gain(%)
Platform Comparison / Usage Preference	0.5832	0.673	15.4
Low Value for Money	0.6439	0.714	10.89
Customer Support Failures	0.603	0.7038	16.72
Login & Access Issues	0.7637	0.7842	2.68
Shipping & Fee Complaints	0.5825	0.6785	16.48
Negative Recommendation	0.7416	0.7592	2.37
Low Ratings & Fairness Concerns	0.8212	0.8239	0.33

Table 4 presents the comparative results across seven aspects. The table includes three columns: *Avg_Cos.Sim.*, *Consensus*, and *Consensus Gain (%)*. The *Avg_Cos.Sim.* column reports the average cosine similarity between all sentences and the aspect centroid in the baseline, which includes all sentences without any filtering. The *Consensus* column shows the average similarity among only those sentences that meet or exceed the threshold of 0.70. Both *Avg_Cos.Sim.* and *Consensus* scores reflect the degree of internal semantic alignment within each cluster. The *Consensus Gain (%)* column indicates the relative improvement, calculated as the percentage increase from the baseline similarity to the LGDM-based consensus score. For example, in the Login & Access Issues aspect, the baseline similarity of 0.7637 increased to 0.7842 after applying the similarity threshold, yielding a positive consensus gain of +0.0205. A sentence such as "I couldn't log in with my phone number and never received a verification code" is likely to have high similarity to the aspect centroid and thus be retained. In contrast, a sentence like "I deleted the app because it kept sending me promotional emails," while still reflecting dissatisfaction, may fall below the threshold due to weaker semantic alignment with the core login-related theme. These differences are also reflected in Figure 4, which visualizes the cosine similarity distributions of sentences across aspects under different similarity thresholds. In aspects such as *Customer Support Failures*, *Shipping & Fee Complaints*, and *Platform Comparison*, a substantial proportion of sentences fall below the 0.70 threshold indicating higher semantic fragmentation. Conversely, aspects like *Login & Access Issues* and *Negative Recommendation* display tighter distributions clustered above the threshold, supporting the observed consensus gains. The figure illustrates how the degree of semantic concentration varies by aspect, aligning with the improvements captured by the consensus scores.

The results highlight the capability of the LGDM framework to filter out semantically off-topic or noisy content, thereby improving the interpretability, coherence, and diagnostic clarity of aspect-level clusters. Additionally, the framework is effective in revealing aspects with inherently diffuse or fragmented concerns, as evidenced by lower consensus scores in certain categories. This dual functionality refining semantically cohesive topics while surfacing areas of fragmentation positions the approach as a valuable tool for both summarization and platform-level issue diagnosis.

5. CONCLUSION

User generated content, especially online reviews are a valuable source of information for customer insights and natural language processing techniques are essential for extracting and analyzing this information. In particular low rated reviews often contain complaints that can help to identify service failures and improve customer satisfaction. Advanced Large Language Models (LLMs) such as BERT, RoBERTa, GPT-based and their derivatives like Sentence-BERT (SBERT), have emerged as powerful tools for analyzing textual data. They are highly effective in tasks such as aspect detection, identifying implicit sentiment, sarcasm and specific areas of dissatisfaction with their ability to model deep contextual relationships and generate high quality sentence embedding. Integrating these language models into review analysis frameworks enhances analytical precision and also aligns with the growing trend toward intelligent language aware systems in decision making process.

This study focused on analyzing user complaints by examining low rated Turkish language reviews from a second-hand marketplace mobile application available on the Google Play Store. A fine-grained analysis was conducted by integrating sentence level semantic modeling with Large Group Decision Making (LGDM). Unlike traditional Aspect Based Sentiment Analysis (ABSA) approaches that focus primarily on identifying aspect terms and classifying their associated sentiment, the proposed method uses a semantic consensus threshold to filter out weakly aligned sentences. By employing SBERT-based sentence embeddings, it captures deeper contextual nuances in user feedback. Complaint themes are extracted through a two-stage clustering process and labeled using TF-IDF based aspect identification, ensuring that only semantically coherent and widely shared concerns are retained.

To measure the degree of user agreement within each aspect, a cosine similarity based consensus metric was applied to sentence embeddings. The results revealed varying levels of consensus across complaint categories, highlighting both widely shared systemic issues and more fragmented concerns. This consensus aware framework offers actionable insights for platform managers by identifying which issues are most urgently shared among users and warrant immediate attention.

Findings indicate that users exhibit the highest semantic consensus on aspects such as Low Ratings & Fairness Concerns, Login & Access Issues, and Negative Recommendations. In contrast, aspects like Shipping & Fee Complaints and Platform Comparison elicited more diverse opinions, suggesting less unified dissatisfaction in those areas.

This research extends the application of LGDM to a previously underexplored domain second hand marketplace platforms. It demonstrates how consensus modeling can be adapted to informal, large scale, user generated content and offers a scalable, interpretable method for aggregating and prioritizing complaints in digital consumer environments.

Future work can build on this foundation by exploring platform specific patterns of agreement or disagreement. This can be achieved by collecting customer reviews from different second-hand platforms and comparing them to examine how consensus on complaints varies across platforms. Such an analysis would uncover both platform dependent and global service issues, offering insights for targeted and universal improvement strategies.

Author Contribution

Ahmet Cumhur ÖZTÜRK contributed to all stages of the study.

Conflict of Interest

The author has declared no conflict of interest.

REFERENCES

[1] A. Dimoka, Y. Hong, and P. A. Pavlou, "On product uncertainty in online markets: Theory and evidence," MIS Q., vol. 36, no. 2, pp. 395–426, 2012.

[2] N. Hajli, "The impact of positive valence and negative valence on social commerce purchase intention," Inf. Technol. People, vol. 33, no. 2, pp. 774–791, 2020.

[3] X. Yuan, T. Xu, S. He, and C. Zhang, "An online review data-driven fuzzy large-scale group decision-making method based on dual fine-tuning," Electronics, vol. 13, no. 14, p. 2702, 2024.

[4] S. P. Ladella, S. Joginpally, and K. N. Ranjit, "Aspect-Based Sentiment Analysis for Online Products," Journal of Marketing Development & Competitiveness, vol. 18, no. 4, 2024.

[5] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," Int. J. Electron. Commer., vol. 17, no. 2, pp. 99–126, 2012.

[6] O. A. El-Said, "Impact of online reviews on hotel booking intention: The moderating role of brand image, star category, and price," Tour. Manag. Perspect., vol. 33, p. 100604, 2020.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," arXiv preprint arXiv:1908.10084, 2019.

[8] I. P. Carrascosa, Large Group Decision Making: Creating Decision Support Approaches at Scale. Cham, Switzerland: Springer, 2018.

[9] F. Ji, Q. Cao, H. Li, H. Fujita, C. Liang, and J. Wu, "An online reviews-driven large-scale group decision making approach for evaluating user satisfaction of sharing accommodation," Expert Syst. Appl., vol. 213, p. 118875, 2023, doi: 10.1016/j.eswa.2022.118875.

[10] X. Wu, H. Liao, and M. Tang, "Product ranking through fusing the wisdom of consumers extracted from online reviews on multiple platforms," Knowl.-Based Syst., vol. 284, p. 111275, 2024.

[11] W. Ma, F. Ji, C. Liang, Q. Sun, and J. Wu, "A deep learning and large group consensus based cruise satisfaction evaluation model with online reviews," Inf. Sci., vol. 676, p. 120801, 2024.

[12] J. Shi, Y. Zhang, H. Liu, and M. Chen, "Evaluating cruise user satisfaction through online reviews: A method based on sentiment analysis and large-scale group decision-making," Appl. Intell., vol. 55, no. 6, pp. 418–432, 2025.

[13] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manag., vol. 24, no. 5, pp. 513–523, 1988.

[14] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in Adv. Neural Inf. Process. Syst., vol. 28, 2015.

[15] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," arXiv preprint arXiv:1705.02364, 2017.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., vol. 1, pp. 4171–4186, Jun. 2019.

[17] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," arXiv preprint arXiv:2104.08821, 2021.

[18] P. Sircar, A. Chakrabarti, D. Gupta, and A. Majumdar, "Distantly supervised aspect clustering and naming for e-commerce reviews," in Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. - Ind. Track, pp. 94–102, 2022.

[19] R. Saha, "Influence of various text embeddings on clustering performance in NLP," arXiv preprint arXiv:2305.03144, 2023.

[20] Y. J. Zhou, M. Zhou, J. B. Yang, B. Y. Cheng, and J. Wu, "Decentralized multipartite consensus model for multi-attribute group decision making: A user experience-oriented perspective," Expert Syst. Appl., p. 127917, 2025.

[21] I. UI Haq, M. Pifarré, and E. Fraca, "Novelty evaluation using sentence embedding models in open-ended cocreative problem-solving," Int. J. Artif. Intell. Educ., pp. 1–28, 2024.

[22] S. Jin and M. Tsujimoto, "Towards trust building and sustainability on second-hand platforms: A study of Mercari in Japan," Journal of Cleaner Production, vol. 503, Art. no. 145237, 2025.

[23] P. Fors, A. Nuur, and F. Randia, "Conceptualising the peer-to-peer second-hand practice-as-entity," Cleaner and Responsible Consumption, vol. 9, p. 100119, 2023.

[24] Y. Ortakci and B. Borhan, "Optimizing SBERT for long text clustering: two novel approaches with empirical insights," The Journal of Supercomputing, vol. 81, no. 8, p. 950, 2025.

[25] A. A. Akın and M. D. Akın, "Zemberek, an open source NLP framework for Turkic languages," in Proc. 3rd Int. Balkan Conf. Commun. Netw. (BalkanCom), Istanbul, Turkey, 2007.

[26] E. Bigne, C. Ruiz, C. Perez-Cabañero, and A. Cuenca, "Are customer star ratings and sentiments aligned? A deep learning study of the customer service experience in tourism destinations," Service Business, vol. 17, no. 1, pp. 281–314, 2023.

[27] D. García-Zamora, E. Herrera-Viedma, F. J. Cabrerizo, J. Liu, and W. Pedrycz, "Large-scale group decision making: A systematic review and a critical analysis," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 6, pp. 949–966, Jun. 2022.