Urban Company Service Review Analysis

BATTALA DIVYATEJA (2021MCS2126)*, LALIT MEENA(2019CS50439)*, and LAKSHAY KAKKAR(2023CSY7548)*, IIT Delhi, India

1 INTRODUCTION

In the rapidly evolving landscape of digital services, a plethora of online platforms has emerged, transcending traditional boundaries by not only facilitating the delivery of goods but also by orchestrating seamless access to a spectrum of services. These platforms, exemplified by industry leaders such as Urban Company, HouseJoy, Zimmber, Angi, and others, have revolutionized how professionals connect with customers, offering a convenient and efficient channel for service scheduling.

With the surge in digital interactions, customer reviews have become instrumental in shaping perceptions of these platforms and the professionals associated with them. Platforms like Urban Company have garnered extensive feedback on the Google Play Store and other review sites, where users share their experiences regarding both the platform's functionalities and the quality of services provided by professionals.

This project embarks on a comprehensive exploration of diverse review datasets, aiming to extract valuable insights and categorize reviews into distinct segments, notably differentiating between "platform-specific" and "service-specific" feedback. Our primary dataset is sourced from Urban Company's reviews on the Google Play Store, hereinafter referred to as the 'online dataset' or 'Play Store dataset.' To complement our analysis, we have been granted access to a private dataset, further enriching our understanding of customer sentiments and experiences.

Adopting a multifaceted approach, we employ advanced methodologies such as Latent Dirichlet Allocation (LDA), BERTopic, and the Large Language Model (LLM) – Meta's Llama 13B. These techniques are meticulously applied to unravel patterns, topics, and sentiments embedded in the reviews, providing a nuanced understanding of the customer discourse.

The focal point of our study revolves around a comparative analysis between outcomes derived from the analysis of the online dataset and the private dataset. This comparative lens not only sheds light on subtle differences but also unlocks valuable insights into the intricate dynamics of customer feedback, contributing to a holistic understanding of the online service ecosystem.

2 LITERATURE WORK

The literature work in the area of service review analysis for companies like Urban Company is still relatively unexplored, with limited research and available datasets. However, some pertinent studies and tools have laid the groundwork for understanding and analyzing customer feedback in this context.

(1) **An Empirical Study on Urban Company Service Quality Influence on Customers, Hyderabad** by Pravena and Bharathi (2022) delves into the service quality provided by Urban Company in Hyderabad. The study utilizes a questionnaire as a research instrument to analyze the influence of service quality on customers. The

*All authors contributed equally to this article.

Authors' address: Battala Divyateja (2021MCS2126), mcs212126@cse.iitd.ac.in; Lalit Meena(2019CS50439), cs5190439@iitd.ac.in; Lakshay Kakkar(2023CSY7548), csy237548@cse.iitd.ac.in, IIT Delhi , Hauz Khas , New Delhi, Delhi, India, 110016.

findings highlight the significance of online home-based services and the positive impact of Urban Company's quality services on customers' lives.

(2) **BERTopic: Neural topic modeling with a class-based TF-IDF procedure** by Grootendorst (2022) introduces BERTopic, a topic modeling technique that leverages BERT embeddings and class-based TF-IDF for creating dense clusters. This approach enhances the interpretability of topics while maintaining important words in topic descriptions. BERTopic is applied to generate coherent topics, demonstrating competitiveness across various benchmarks.

(3) **Latent Dirichlet Allocation (LDA)** is a generative probabilistic model for discrete data, often used in natural language processing and topic modeling. LDA explains observations through unobserved groups, revealing latent topics in documents. It is based on probability distributions, assuming that documents with similar topics will use a similar group of words. LDA aims to find a lower dimensionality representation of data while preserving its statistical structure.

(4) The project incorporates the **google-play-scraper** library's API, a tool for scraping reviews from the Google Play Store. This library facilitates the collection of valuable data for analysis, contributing to a more comprehensive understanding of customer sentiments and experiences related to online platforms and services.

3 ANALYSIS OF DATASET

We have done comparative analysis of URBANCOMPANY private data with the online reviews dataset to draw out interesting insights from these dataset. Here is the comparision table (**Table 1**)-

Table 1. Dataset Information

Online Playstore Dataset	**Private Dataset**
The dataset consists of 107,062 reviews with 11 features.	The dataset consists of 1,554,516 records with 10 features.
Only 4 reviews are blank.	Private Dataset also includes information about the reviewer's city.
App, urban, services are the most frequently occurring words in the dataset.	After preprocessing, we are only left with 375,869 records, implying most of the reviews are blank. Additionally, there are also some blank reviews without any ratings. Work, Professional, Job, Great are some of the most occurring words in the private dataset.

Here below is the Frequent words comparison of these datasets (**Fig. 1**)-

3.1 Analysis of Online Reviews Data

Reviews of Urbancompany app were scraped from Google Play Store using *google-play-scraper* library as mentioned in the above section. The dataset consists of 1,07,062 records containing information about the *review content, rating, reply by app publisher* and so on. The analysis of the dataset follows: The Word Cloud of the reviews data is shown in **Fig. 2**.

The histogram plot of *Token Length* vs *Frequency* is depicted in **Fig. 3**.

The following are the other statistics related to length of the reviews:

Mean: 16.4
Median: 5.0

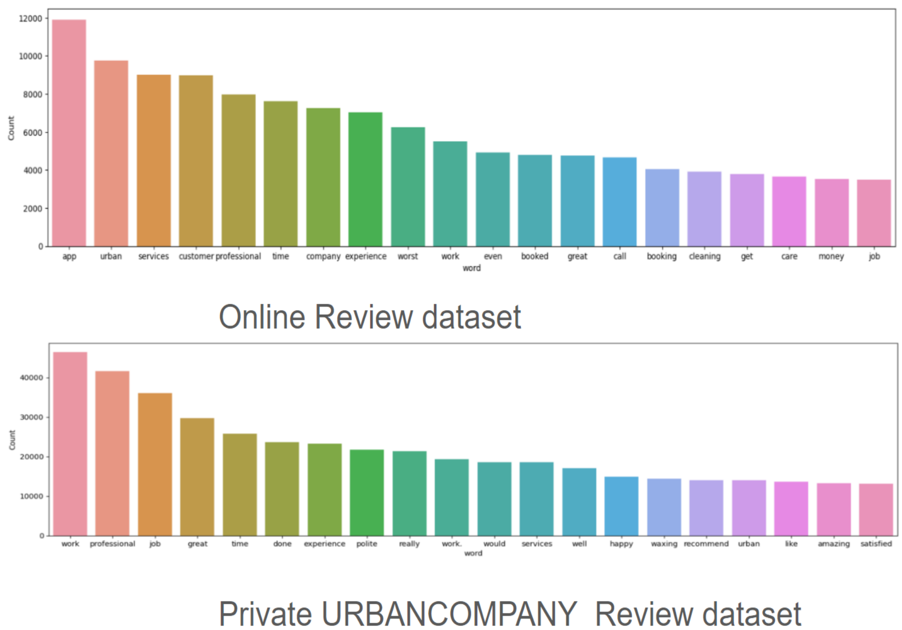


Fig. 1. Frequent words comparison of datasets involved

Fig. 2. Word Cloud of the online reviews data

Standard Deviation: 23.2
Variance: 540.2
Minimum Value: 1
Maximum Value: 492

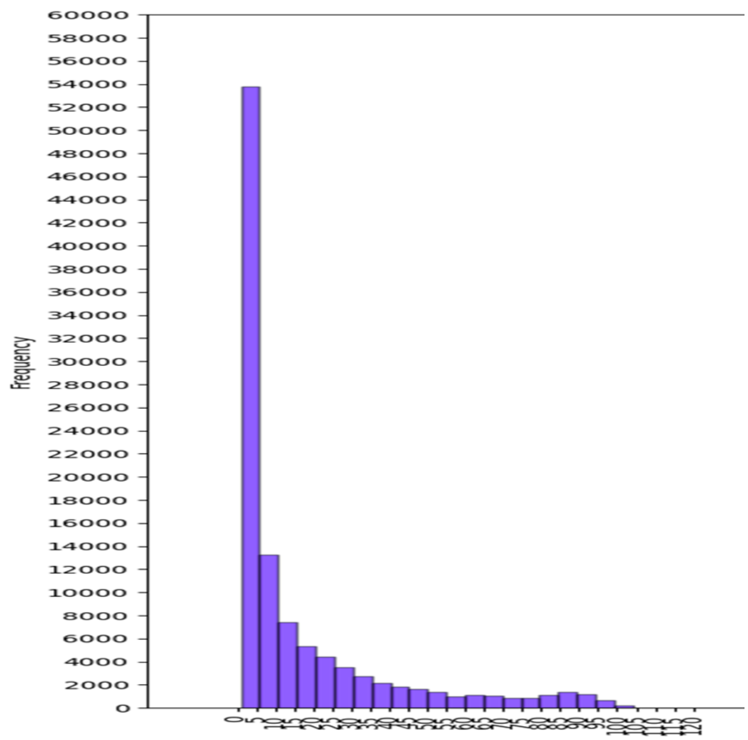


Fig. 3. Histogram plot: *Token Length* vs *Frequency*

We have also done Analysis of Rating Scores of these online reviews data so to correlate length of reviews and given rating score by user for any useful insight (**Fig. 4**)

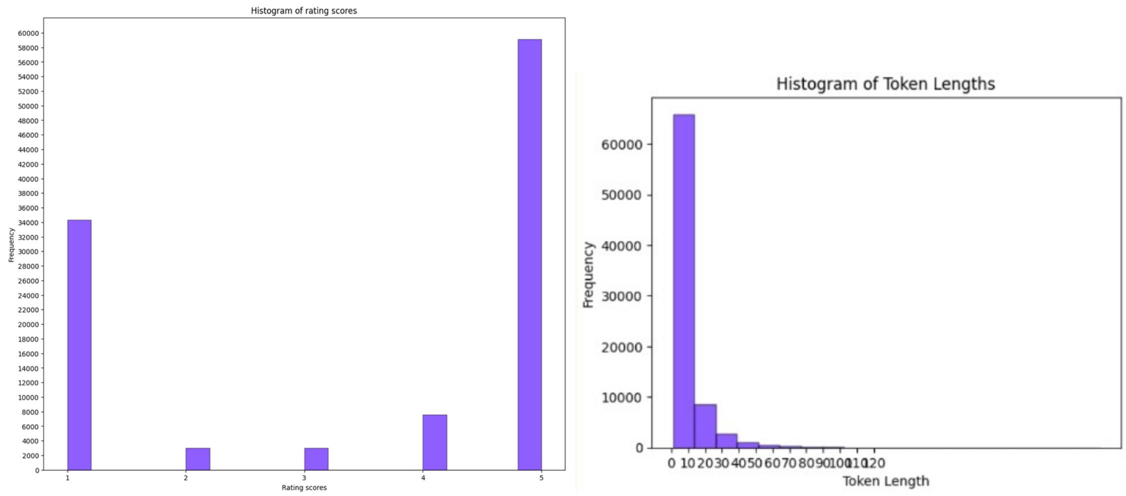
From this **review length vs score analysis**, a key point to note here, is that good reviews (ones which carry 5 star) tend to be short in length whereas bad reviews tend to be more descriptive.

4 METHODOLOGY

We have used three different methods for identifying the reviews in the original dataset(play store dataset) using the private dataset given. The first two methods used the LDA (Linear Discriminant Analysis) and Latent space approach for identifying the reviews present in the original dataset.
Here is the list of Results with our experimental Analysis-

4.1 Baseline LDA Method

The model starts with the preprocessing step in which the techniques such as lematization, stop word removal and normalization. Then the LDA method is used to identify the topics present in the original dataset. The number of topics is set as 10 after doing thorough analysis. Then the topics identified are assigned to the service, platform and Both labels. Then created the LDA topic vectors for each review present in the original dataset. Then calculated the topic distributions for each review and based on the threshold, they are classified as service, platform or both. The threshold



(a) ANALYSIS OF RATING SCORES (b) ANALYSIS OF REVIEW LENGTH of 5 STAR RATING SCORES

Fig. 4. Analysis of Rating Scores to correlate it with length of reviews

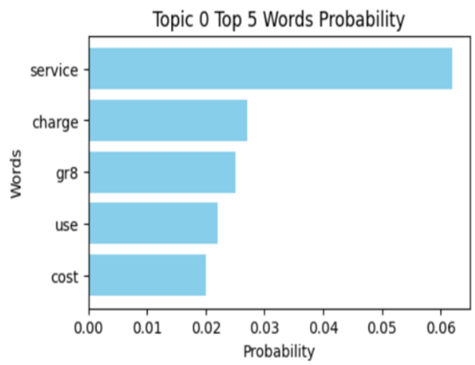
is taken as 0.5. Based on the threshold, calculated the propobality distribution for the topics of the service, platform and both for each review. If the sum of their probability distribution is greater than 0.5 , considered as service, platform or both. Then the reviews are verified with the set of 100 examples for which manual annotation of the reviews is done.

Topics Identified using LDA

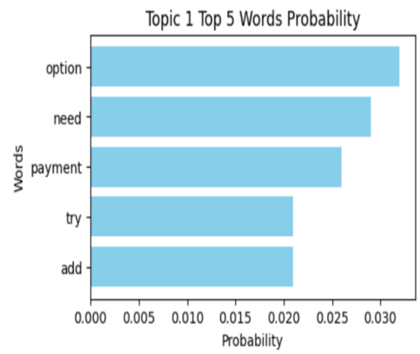
```
[(0,
  '0.379*"service" + 0.062*"charge" + 0.027*"gr8" + 0.025*"use" + 0.022*"cost" + 0.020*"lot" + 0.018*"market" + 0.018*"price" +
0.017*"become" + 0.015*"base"'),
 (1,
  '0.042*"option" + 0.032*"need" + 0.029*"payment" + 0.026*"try" + 0.021*"add" + 0.021*"make" + 0.021*"aap" + 0.020*"aur" + 0.01
7*"offer" + 0.016*"find"'),
 (2,
  '0.077*"company" + 0.033*"clap" + 0.029*"need" + 0.027*"guy" + 0.023*"life" + 0.021*"solution" + 0.021*"technician" + 0.021*"t
ake" + 0.019*"place" + 0.018*"bill"'),
 (3,
  '0.087*"gud" + 0.076*"application" + 0.042*"money" + 0.026*"refund" + 0.025*"chor" + 0.024*"balance" + 0.023*"t" + 0.021*"pay"
+ 0.017*"prompt" + 0.016*"waste"'),
 (4,
  '0.403*"app" + 0.116*"love" + 0.029*"use" + 0.023*"awsome" + 0.019*"user" + 0.012*"area" + 0.009*"machine" + 0.009*"location"
+ 0.006*"washing" + 0.006*"address"'),
 (5,
  '0.115*"customer" + 0.052*"response" + 0.038*"care" + 0.025*"team" + 0.021*"support" + 0.018*"call" + 0.014*"number" + 0.014
*"worker" + 0.013*"contact" + 0.012*"way"'),
 (6,
  '0.292*"experience" + 0.185*"work" + 0.154*"job" + 0.032*"thank" + 0.031*"improve" + 0.019*"do" + 0.013*"recommend" + 0.012*"c
arpenter" + 0.011*"clean" + 0.009*"cook"'),
 (7,
  '0.054*"professional" + 0.045*"star" + 0.045*"people" + 0.029*"home" + 0.025*"costlier" + 0.025*"quality" + 0.025*"give" + 0.0
22*"massage" + 0.018*"provider" + 0.018*"urbanclap"'),
 (8,
  '0.046*"help" + 0.039*"get" + 0.029*"problem" + 0.028*"issue" + 0.023*"day" + 0.022*"repair" + 0.021*"call" + 0.020*"gas" + 0.
019*"year" + 0.017*"fix"'),
 (9,
  '0.088*"time" + 0.045*"book" + 0.037*"bakwa" + 0.027*"day" + 0.025*"call" + 0.024*"say" + 0.022*"hour" + 0.020*"request" + 0.0
19*"schedule" + 0.018*"show"')]
```

Fig. 5. Lda topics identified stats

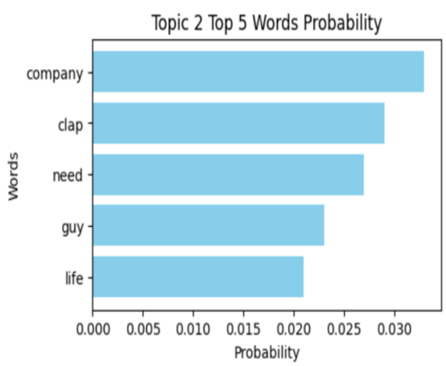
Topics along with top words probability distributions These are the topics identified along with the words with their probability distribution of words



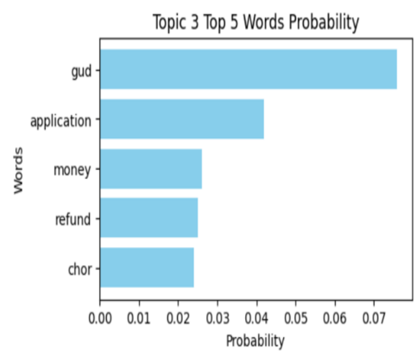
(a) Topic0 words probability distribution



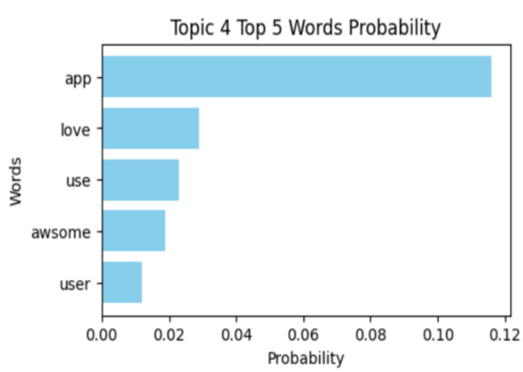
(b) Topic1 words probability distribution



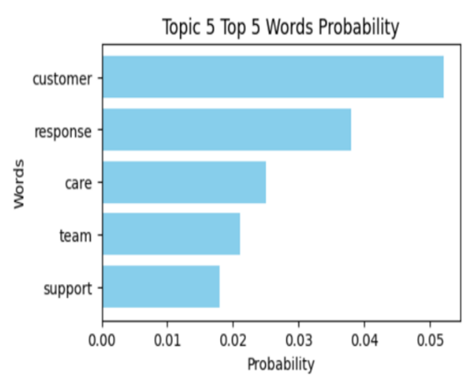
(a) Topic2 words probability distribution



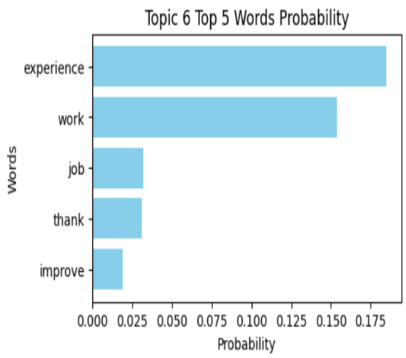
(b) Topic3 words probability distribution



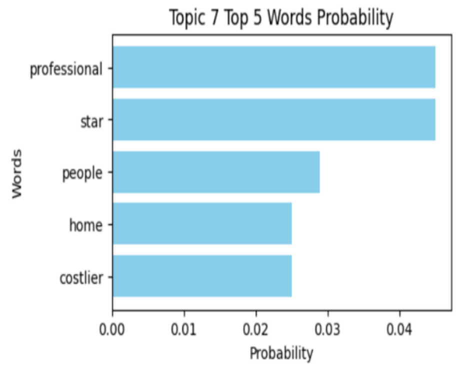
(a) Topic4 words probability distribution



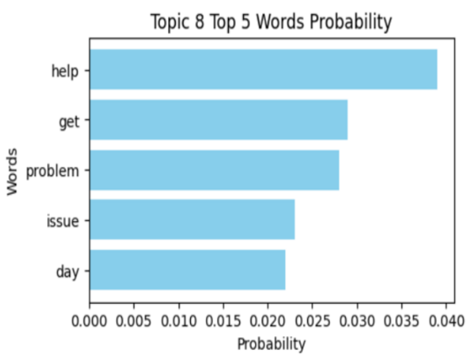
(b) Topic5 words probability distribution



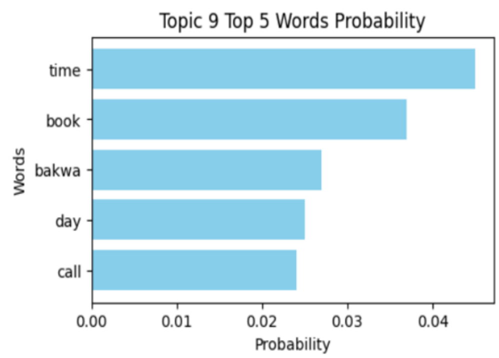
(a) Topic6 words probability distribution



(b) Topic7 words probability distribution



(a) Topic8 words probability distribution



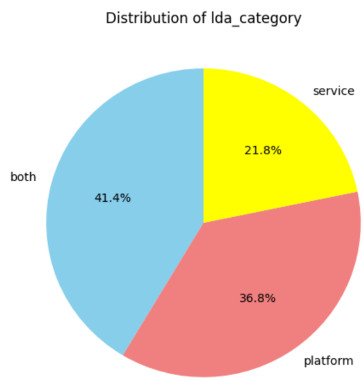
(b) Topic9 words probability distribution

Results:

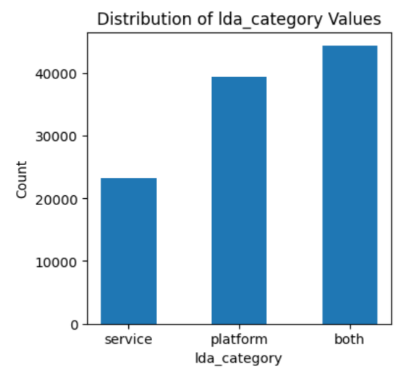
content	lda_category
There clarity for the services and there tarrifs. You opt for service you get quote from the service guy that you accept pay. When you get the tax invoice, different from the one you accepted first place. Then you ask for revised correct invoice, you are left with vague responses and explanations. can't even get the correct invoice the service taken from UC. Not done. Disappointed!!	Platform
Hi Shivesh,We're really sorry that you're experiencing problems. You can always reach out to us through the help center section on our app. However, please help us with your registered contact details and concerns at resolve urbancompany.com and we will have a call arranged for you.	Platform
Hi Shatabhisha,We apologize for the inconvenience caused. Kindly share your contact details with us at resolve urbancompany.com. We will prioritize your concern and reach out to you via call to address the issue promptly.Regards, Team Urban Company	Platform
Hi Madhurjya, Apologies for the less-than-ideal experience. We request you to please share your registered mobile number over resolve urbancompany.com so that we can help you in a better way. Regards, Team Urban Company	Platform
Hi Navlyn, We have taken your feedback and it will give us a chance to improve our services further. We look forward to serve you better in the future! Regards, Team Urban Company	Service

Fig. 11. Lda topics identified

The Resultant distribution of the reviews in the dataset are given below:



(a) Distribution of Reviews

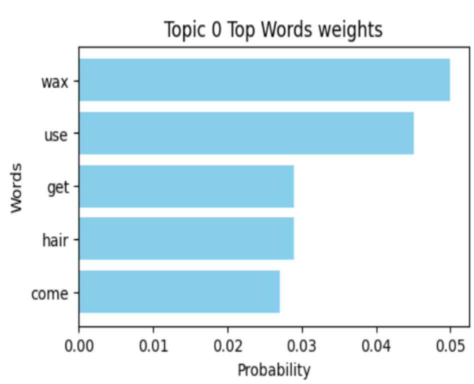


(b) Distribution of Reviews

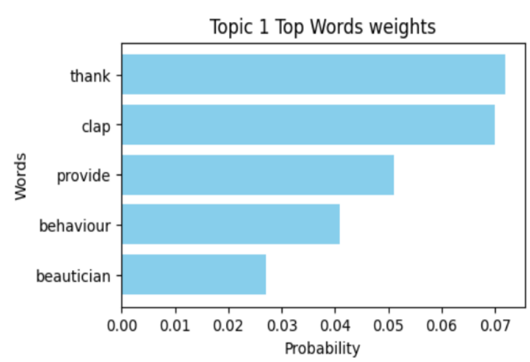
4.2 Latent Space Method

The model starts with similar preprocessing step in which the techniques such as lematization, stop word removal and normalization. Then applied the latent space technique to create the latent space on the private dataset. Then applied the same technique on the original dataset and based on the distance between the latent space and the topic vector of the each review, classified the reviews as service and platform specific.

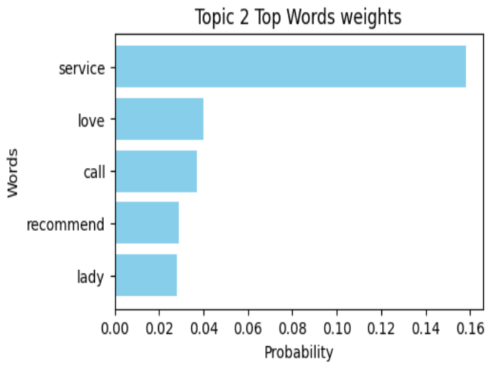
Topics along with top words probability distributions of the original dataset These are the topics identified along with the words with their probability distribution of words



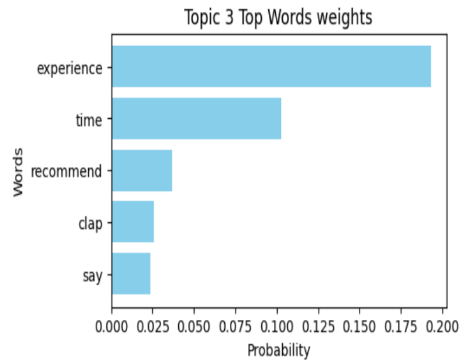
(a) Topic0 Probability distribution



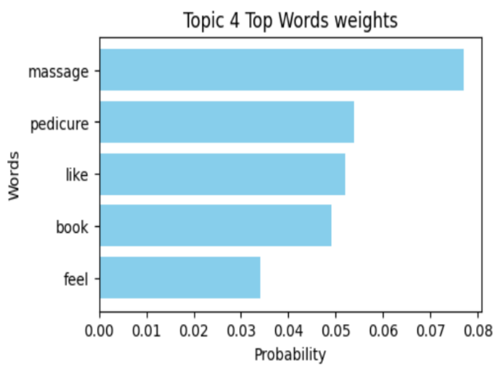
(b) Topic1 Probability distribution



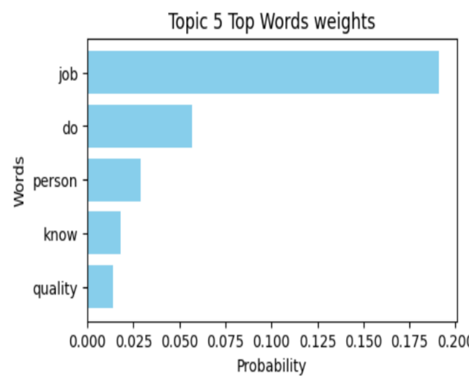
(a) Topic2 Probability distribution



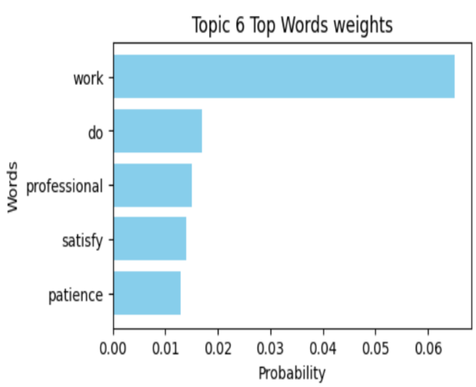
(b) Topic3 Probability distribution



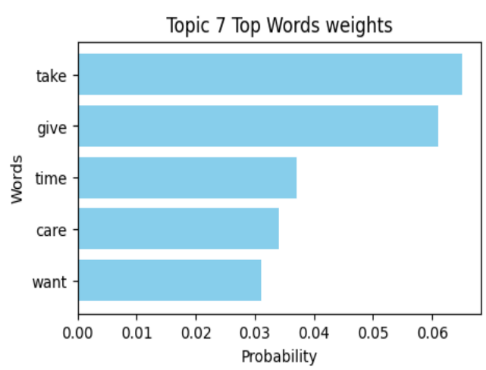
(a) Topic4 Probability distribution



(b) Topic5 Probability distribution



(a) Topic6 Probability distribution



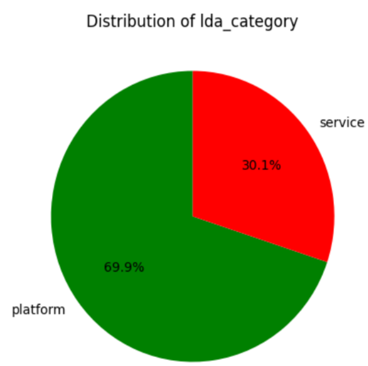
(b) Topic7 Probability distribution

Results:

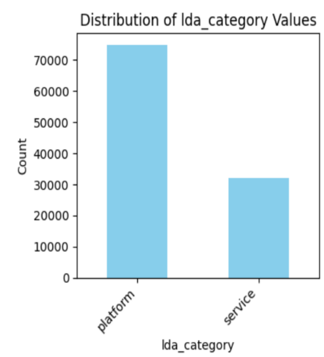
content	lda_category
There clarity for the services and there tarrifs. You opt for service you get quote from the service guy that you accept pay. When you get the tax invoice, different from the one you accepted first place. Then you ask for revised correct invoice, you are left with vague responses and explanations. can't even get the correct invoice the service taken from UC. Not done. Disappointed!!	Platform
Hi Shivesh,We're really sorry that you're experiencing problems. You can always reach out to us through the help center section on our app. However, please help us with your registered contact details and concerns at resolve urbancompany.com and we will have a call arranged for you.	Platform
Hi Shatabhisha,We apologize for the inconvenience caused. Kindly share your contact details with us at resolve urbancompany.com. We will prioritize your concern and reach out to you via call to address the issue promptly.Regards, Team Urban Company	Platform
Hi Madhurjya, Apologies for the less-than-ideal experience. We request you to please share your registered mobile number over resolve urbancompany.com so that we can help you in a better way. Regards, Team Urban Company	Platform
Hi Navlyn, We have taken your feedback and it will give us a chance to improve our services further. We look forward to serve you better in the future! Regards, Team Urban Company	Service

Fig. 17. Lda topics identified

The distribution of the reviews in the dataset are given below:



(a) Distribution of Reviews



(b) Distribution of Reviews

4.2.1 ***Analysis and Verification of LDA and Latent Space Methods.*** The results of the LDA method and latent space method are verified using a sample set of 100 examples where manual annotation of the labels is done.

The results for the LDA method are given below:

Total Reviews: 100
Correct Predictions: 68
Wrong Predictions: 32
Total Platform Reviews: 74
Total Service Reviews: 18
Both: 8
Accuracy: 0.68

The results for the latent space method are given below:

Total Reviews: 100
Correct Predictions: 73
Wrong Predictions: 27
Total Platform Reviews: 80
Total Service Reviews: 20
Accuracy: 0.73

In the results, it is evident that both approaches predict a higher number of platform-specific reviews, aligning with the manual annotation, which indicates a prevalence of platform-specific feedback. Notably, the updated Latent Space method exhibits a distinct improvement over the previous LDA method. It effectively distinguishes platform-specific reviews, showcasing enhanced performance in this regard.

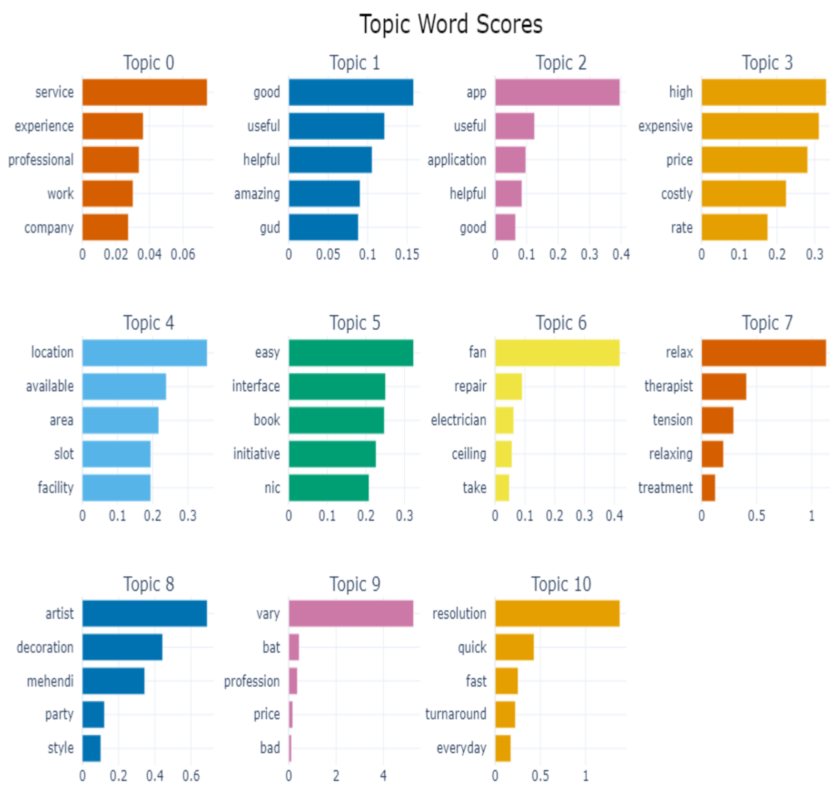


Fig. 19. Extracted topics from BERTopic model

4.3 BERTopic

BERTopic is a topic modeling python library that combines transformer embeddings and clustering model algorithms to identify topics in NLP (Natual Language Processing). BERTopic is a topic modeling technique that leverages BERT embeddings and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.

4.3.1 ***Implementation Details.*** We implemented the topic model using BERTopic. The preprocessing phase involved removing examples with missing reviews, stopwords removal, lemmatization, etc. The model was configured with 11 topics. Relevant topics were identified along with the corresponding probabilities for each document (**Fig. 19**). **Hierarchical clustering** (**Fig. 20**) and the **intertopic distance map**(**Fig. 21**) were employed to map these topics into three categories, leveraging the weighted probability score. The categorization is as follows:

- **Service:** [0, 6, 7, 8, 9]
- **Platform:** [2, 3, 4, 5, 10]
- **Mix:** [1]

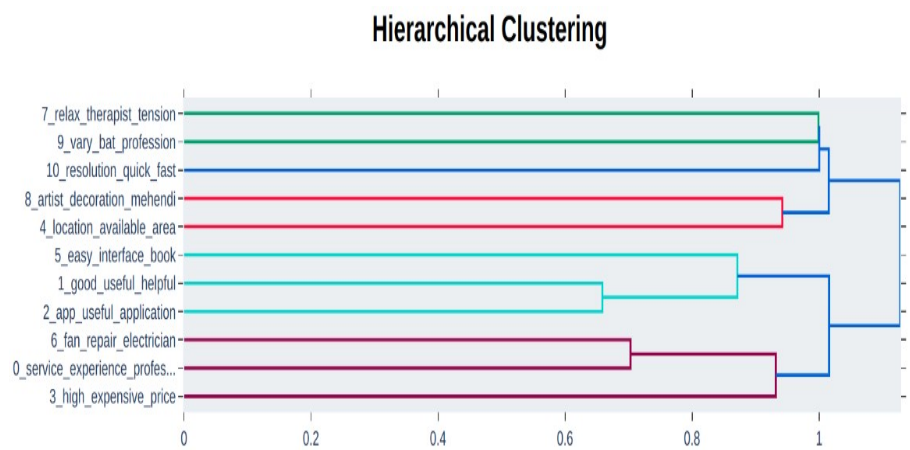


Fig. 20. Dendogram showing hierarchical clustering of topics from BERTopic

4.3.2 ***Results and analysis.*** We have also used Heatmap (**Fig. 22**) to analyze the similarities between topics. The similarity score ranges from 0 to 1. A value close to 1 represents a higher similarity between the two topics, which is represented by darker blue color.We can see that there is high similarity between topics like (2_app_useful_application, 5_easy_interface_book), (6_fan_repair_electrician, 0_service_experience_professional).

Here is also an example of prediction by this method(**Fig. 23**). Output of this example is given in the figure for this given Input -'horrible experience.tried book app payment deduct booking slot reflect help customer service connect app escalate concern unusual lose money service stay app pure case cheat pathetic customer experience basic feature recommend'

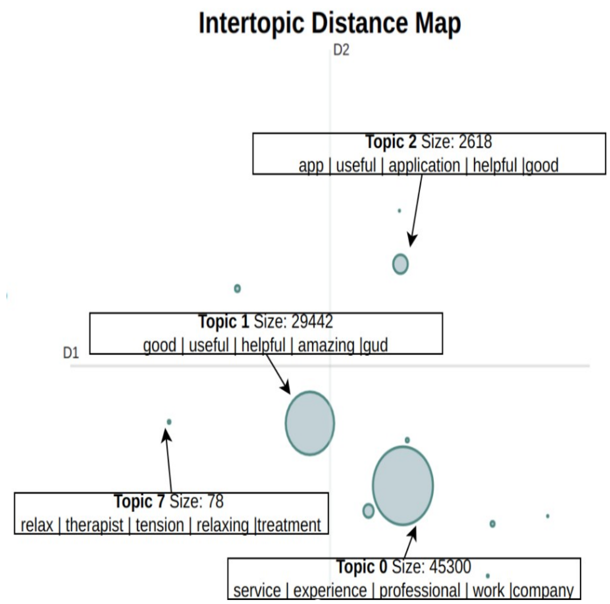


Fig. 21. Intertopic distance map

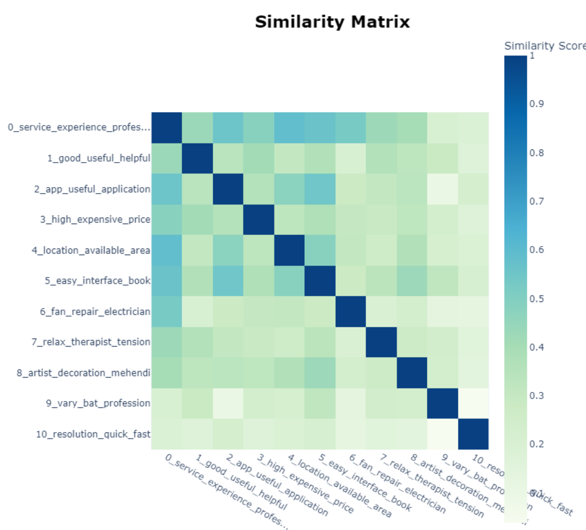


Fig. 22. Heatmap to analyze the similarities between topics

The pie chart below (**Fig. 24**) illustrates the distribution of service-specific and platform-specific reviews within the Online Reviews dataset. Notably, there is a higher prevalence of service-specific reviews compared to platform-specific

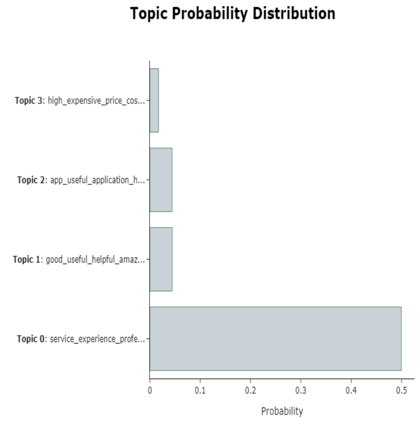


Fig. 23. An example's prediction by BERTopic

reviews. This disparity can be attributed to certain limitations inherent in the NLP-based embedding method employed. A notable challenge is the presence of false positives, particularly arising from complex reviews that encompass both platform-specific and service-specific feedback. Categories such as customer support service, pricing versus service value, etc., contribute to the complexity. To enhance the efficacy of the method in segregating reviews, the inclusion of additional categories related to the company, such as availability, scheduling issues, pricing, booking with the same agent, etc., will be useful in building robust model.

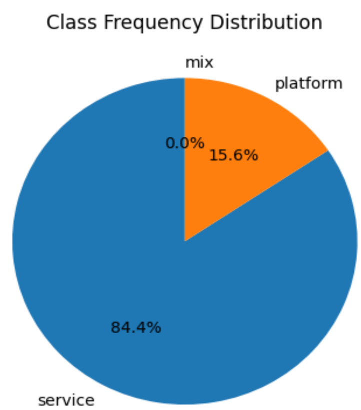


Fig. 24. Class Frequency distribution of BERTopic predictions

4.4 Large Language Model(LLM) based method

In our project, we leveraged Meta's powerful 13B parameter LlaMA language model as our Large Language Model (LLM) of choice. Given the inherent capabilities of LLMs in comprehending language and handling various natural language processing (NLP) tasks effectively, we utilized this model to address the task at hand. The initial step involved presenting the LLM with a well-structured prompt, comprising a detailed problem statement, relevant context, and illustrative examples. This thoughtful input served to guide the LLM in understanding the nuances of the reviews, facilitating a clear segregation between platform-specific and service-specific categories. We instructed the LLM to generate output in two parts: firstly, providing the essential keywords indicative of the review's content, followed by assigning an appropriate label based on whether it pertains to the platform or the service. This sequential approach allowed for a comprehensive and nuanced analysis of the reviews, showcasing the LLM's proficiency in extracting key information and classifying reviews effectively.

4.4.1 ***Structure of given prompt and examples.*** In our project, the prompt structure (**Fig. 25**) played a pivotal role in guiding Meta's formidable 13B parameter LlaMA language model through the nuanced task of classifying customer reviews. The problem was initially presented with a detailed problem description, providing clarity on the overarching challenge at hand. Subsequently, the context section was carefully crafted, offering insights into the landscape of online service platforms and the intricacies involved in customer reviews. Leveraging the power of in-context learning, we supplemented the prompt with examples, allowing the LLM to grasp the intricacies of distinguishing platform-specific and service-specific reviews. The input review section contained a diverse set of customer feedback, ensuring that the LLM was exposed to a comprehensive range of scenarios. Finally, in the output prediction section, we directed the LLM to generate responses in two parts: key keywords indicative of the review content, followed by a precise label categorizing the review as either platform-specific or service-specific.

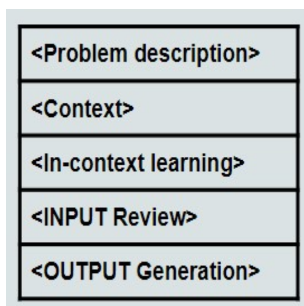


Fig. 25. Structure of given prompt

This strategic structuring of the prompt facilitated a robust and insightful analysis, showcasing the LLM's ability to comprehend complex language patterns and deliver accurate predictions.
Here, we have also mentioned an example corresponding to Service-specific and Platform-specific reviews ((**Fig. 26 and Fi. 27**)-

4.4.2 ***Results analysis.*** To check performance of this LLM based approach, we have taken sample of first 100 reviews which we labeled manually. Then the performance of this LLM based approach was thoroughly assessed using various metrics and visualizations. The confusion matrix, depicted in (**Fig. 28**)below, provides a comprehensive overview of the model's predictions.

```
Result:
Problem is about to classify the following customer reviews as either "Platform-Specific" or "Service-Specific" based on the context of the review. Additionally, extract keywords that correspond to the identified category. For each review, provide a label and a list of relevant keywords.
Context: Some online platforms like Urban Company,Angi etc have started online booking of services. These platforms connect professionals to customers and facilitate the scheduling of services. Platform-Specific reviews are related to app,customer call, pricing and booking  related, the User Interface (UI) is bad). Service-Specific reviews are related to quality of the service, professionalism or their behaviour.
Here are some INPUT-OUTPUT examples for your reference-
INPUT-
Review: App does not open when you need it to open and if it does it is so slow...horrible I have been trying to cancel my booking but it does not happen and the app starts working just when the partner is about to come and then you cannot cancel and is forced to take the service. Pathetic The more number of service you take higher are the taxes without explanation
OUTPUT-
Keywords: [App,booking, taxes]
Label: Platform-Specific
INPUT-
Review: Their 'professionals' are anything but professional. They are impunctual, indisciplined, not reachable, not accountable, not responsible. They switch off the tracking, dont answer the phone, u name it. Its better to get the services from the neighborhood vendors. Theres no way to upload the video here as to how their app and workers work in tandem to fool the customer. Hopeless service.
OUTPUT-
Keywords: [professionals,service]
Label: Service-Specific
INPUT-
Review:I had used the app to install 3 AC's in three different locations and each time I have had issues with technicians. They do not handle the inlet or any product carefully. They always have a irritating answer to our question. If we try complaining customer service you do not have good options to connect directly. Overall technicians aren't trained so well and they lack technical skills. Today I booked a technician and he said that he doesn't want to service and asked me to book again. Horrible
OUTPUT-
Keywords: [technicians, service]
Label: Service-Specific
```

Fig. 26. Example corresponding to Service-specific review

From the matrix, it is evident that the model achieved an accuracy of 65%. The model correctly predicted 40 instances of the **service-specific reviews (positive class)** as compare to only 25 instances of **platform-specific reviews (negative class)**. However, it misclassified more 31 instances of the negative class as positive. This is mostly due to the observation that most of platform specific reviews also contain service-specific feedback and confusing categories like customer support service, pricing vs service value,etc.

Further insights into the model's performance are revealed in the classification report, as shown in Table (**Fig. 29**). Precision, recall, and F1-score metrics are reported for both classes. Notably, the model achieved a low precision of 0.56 for the positive class, indicating a high false-positive rate. The recall and F1-score for the positive class were also respectable, demonstrating the model's ability to capture the majority of positive instances.

Additionally, the Receiver Operating Characteristic (ROC) curve, displayed in (**Fig. 30**), illustrates the trade-off between true positive rate and false positive rate. The area under the ROC curve (AUC) is 0.68, indicating a high discriminatory power of the model.

```
Result:
Problem is about to classify the following customer reviews as either "Platform-Specific" or "Service-Specific" based on the context of the review. Additionally, extract keywords that correspond to the identified category. For each review, provide a label and a list of relevant keywords.
Context: Some online platforms like Urban Company,Angi etc have started online booking of services. These platforms connect professionals to customers and facilitate the scheduling of services. Platform-Specific reviews are related to app,customer call, pricing and booking  related, the User Interface (UI) is bad). Service-Specific reviews are related to quality of the service, professionalism or their behaviour.
Here are some INPUT-OUTPUT examples for your reference-
INPUT-
Review: App does not open when you need it to open and if it does it is so slow...horrible I have been trying to cancel my booking but it does not happen and the app starts working just when the partner is about to come and then you cannot cancel and is forced to take the service. Pathetic The more number of service you take higher are the taxes without explanation
OUTPUT-
Keywords: [App,booking, taxes]
Label: Platform-Specific
INPUT-
Review: Their 'professionals' are anything but professional. They are impunctual, indisciplined, not reachable, not accountable, not responsible. They switch off the tracking, dont answer the phone, u name it. Its better to get the services from the neighborhood vendors. Theres no way to upload the video here as to how their app and workers work in tandem to fool the customer. Hopeless service.
OUTPUT-
Keywords: [professionals,service]
Label: Service-Specific
INPUT-
Review:The application is becoming more and more unusable now a days. Their service persons doesn't show up, the app do not give the option for pay online after service option while booking any service. So they want the money upfront, but without any guarantee that you will get the service on time as promised! Moreover, their non-existent customer support in making the thing worse overall. We need an alternative now. Urban company has lost is completely.
OUTPUT-
Keywords: [application, service, service person, customer support]
Label: Platform-Specific
```

Fig. 27. Example corresponding to Platform-specific review

4.4.3 ***Summary***.

(1) In summary, this approach exhibits promising performance specifically for Service-specific reviews with a notable accuracy, precision, and AUC score.
(2) Some of complicated reviews are those platform specific reviews which contain service-specific feedback also and confusing categories like customer support service, pricing vs service value,etc.
(3) We can improve segregation of reviews by including other important categories related to Company also like availability, scheduling related issues, pricing, Booking with same agent,etc
(4) To improve current model we can make Prompt structure more robust by including more examples,etc
(5) We can also use a more powerful language model with significantly more parameters like GPT-4 for better results.

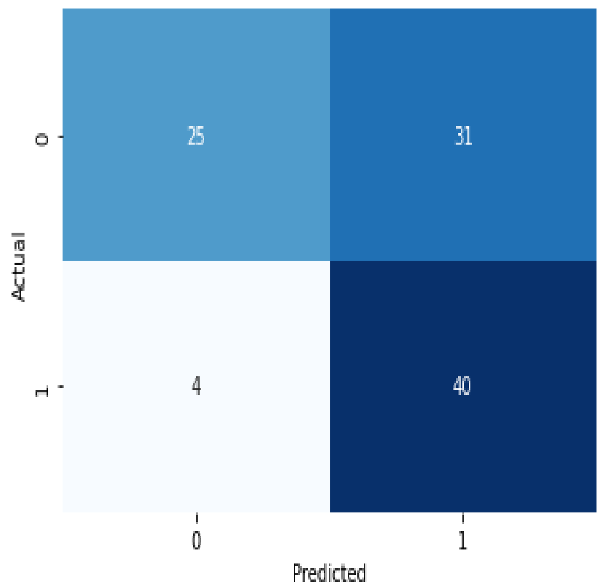


Fig. 28. Fig. Confusion Matrix

	precision	recall	f1-score	support
0	0.86	0.45	0.59	56
1	0.56	0.91	0.70	44
accuracy			0.65	100
macro avg	0.71	0.68	0.64	100
weighted avg	0.73	0.65	0.64	100

Fig. 29. Fig. Classification report

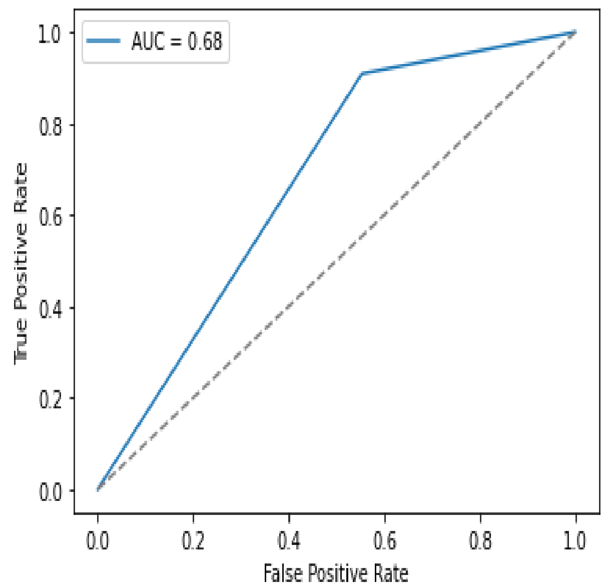


Fig. 30. Fig. Receiver Operating Characteristic (ROC) curve

5 CONCLUSION

In conclusion, our research has yielded valuable insights into the dynamics of online reviews, particularly in the context of platform-specific and service-specific feedback. Through a careful comparison of models and datasets, we observed a prevalence of platform-specific reviews in the online dataset, highlighting the platform-centric nature of technical platforms such as Google Play Store. In contrast, service-specific reviews were more prominent in the private Urban Company dataset, reflecting a distinct focus on the quality and specifics of services.

While all employed methods demonstrated commendable performance, the latent space method exhibited a slight superiority over NLP-based techniques. The comparison revealed that both approaches, LDA method and latent space-based approach, leaned towards predicting a higher number of platform-specific reviews, aligning with manual annotations and emphasizing the prevalence of platform-centric feedback. The enhanced performance of the updated Latent Space method underscores its effectiveness in distinguishing and categorizing platform-specific reviews.

However, NLP and embeddings-based approaches, including the LLM-based method and BERTopic, faced challenges in accurately categorizing certain complex reviews. These reviews, often platform-specific, contained elements of service-specific feedback, leading to confusion in categorization, especially in nuanced areas like customer support service and pricing versus service value.

As part of future work, there is a potential for improvement in review segregation by incorporating additional categories related to the company, such as availability, scheduling issues, pricing, and booking preferences with the same service provider. As these are additional significant topics found in our analysis work. This expansion could contribute to a more nuanced understanding of customer feedback, enhancing the effectiveness of the analysis.

6 REFERENCES

(1) Mrs. S.E. Pravena and Dr. G Bharathi. (2022). *An Empirical Study on Urban Company Service Quality Influence on Customers, Hyderabad*. Retrieved from http://ijmer.in.doi./2022/11.03.97

(2) Maarten Grootendorst. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. Retrieved from https://arxiv.org/abs/2203.05794

(3) Latent Dirichlet Allocation (LDA). Retrieved from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

(4) `google-play-scraper` library. Retrieved from https://github.com/JoMingyu/google-play-scraper