Intelligent Data And Text Analytics.

COURSEWORK 2

UP2225522

Text-Preprocessing

In this report, the data set “yelp_labelled” which is a restaurant review comments are used for text analysis. It is originally a text file but is changed to a CSV using Excel. The data set contains a “Text” column which shows the comments(review) of the restaurant and “Sentiment” which indicates whether the comment is positive(1) or negative(0). Table 1 shows the first 5 rows of the dataset.

Table 1

	Text	Sentiment
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1

Figure 1

Table 2

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Text       1000 non-null   string
 1   Sentiment  1000 non-null   int64
dtypes: int64(1), string(1)
memory usage: 15.8 KB
```

The dataset has 1000 rows and the data type for “Text” is transformed from object to String, Table 2 reflects the change.

Table 3

	Text	Sentiment
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1

0	wow loved this place
1	crust is not good
2	not tasty and the texture was just nasty
3	stopped by during the late may bank holiday of...
4	the selection on the menu was great and so wer...

The data is then pre-processed first, eliminating punctuation marks from the text. All the text is converted to lowercase letters. From Table 3, the left-hand side shows the data before the change and the right is after with yellow labels indicating where the change occurred. Overall, these steps aim to standardize the text for analysis by removing punctuation, and numbers, and ensuring all letters are in the same case (lowercase).

Table 4

0	wow loved this place	0	wow loved place
1	crust is not good	1	crust good
2	not tasty and the texture was just nasty	2	tasty texture nasty
3	stopped by during the late may bank holiday of...	3	stopped late may bank holiday rick steve recom...
4	the selection on the menu was great and so wer...	4	selection menu great prices

Furthermore, the stop words are also removed, Table 4 reflects the changes. This process aims to remove commonly occurring words that may not carry specific meaning for analysis, making the text more focused on meaningful content.

Table 5

0	wow loved place	0	wow love place
1	crust good	1	crust good
2	tasty texture nasty	2	tasty texture nasty
3	stopped late may bank holiday rick steve recom...	3	stop late may bank holiday rick steve recommen...
4	selection menu great prices	4	selection menu great price

The final step involves lemmatization, a process that converts words into their base or dictionary form (known as lemma), considering their part of speech. It transforms words into their most basic form to streamline analysis. For example, "loved" becomes "love," "stopped" becomes "stop," and "prices" remain unchanged as it's already in its base form. This transformation reduces variations of words to their root, facilitating clearer analysis by treating different forms of a word as a single entity. It makes the text more consistent and helps focus on the importance of the words in their simplest, most interpretable forms.

Table 6

	Review	Sentiment
0	wow love place	1
1	crust good	0
2	tasty texture nasty	0
3	stop late may bank holiday rick steve recommen...	1
4	selection menu great price	1
...	...	...
995	think food flavor texture lack	0
996	appetite instantly go	0
997	overall impressed would go back	0
998	whole experience underwhelming think well go n...	0
999	hadnt waste enough life pour salt wound draw t...	0

1000 rows × 2 columns

The data is then merged back with the “Sentiment” column. Table 6 reflects the change.

Figure 2

Table 7

```
  (0, 1095)     0.3775149654879757
  (0, 855)      0.5213223479284713
  (0, 1654)     0.7653139619678703
  (1, 625)      0.44530826402032975
  (1, 342)      0.895377322694293
  (2, 962)      0.6095408468238828
  (2, 1471)     0.6095408468238828
  (2, 1459)     0.5068726783980841
  (3, 1196)     0.3285884923458373
  (3, 1397)     0.3635078590312358
  (3, 1232)     0.3635078590312358
  (3, 704)      0.3635078590312358
  (3, 99)       0.3635078590312358
  (3, 891)      0.3081619722859476
  (3, 813)      0.3430813389713461
  (3, 1404)     0.31734696715413374
  (3, 855)      0.2238303922896418
  (4, 1132)     0.5090497757419697
  (4, 637)      0.385344670765044
  (4, 912)      0.525888742898675
  (4, 1290)     0.5619776157663946
  (5, 1083)     0.4572283769888474
  (5, 353)      0.4830289242429889
  (5, 1599)     0.38919034733963315
  (5, 36)       0.553289705826775
...
  (999, 1261)   0.3000752750794398
  (999, 476)    0.24584953822953573
  (999, 1449)   0.23350766347522658
  (999, 1495)   0.18033475183648912
```

After the text has been pre-processed by removing stop words, lemmatizing words, and preparing it in a cleaned format, the TF-IDF vectorization process is applied.

The values in the output represent the importance of words in each review. Higher values indicate more significance of a word within a particular review. The TF-IDF process boosts the weight of words that are uncommon across all documents but are frequent within a specific document. This helps in identifying words that are important for a particular review but not so common across all the reviews. This numerical representation allows machine learning models to understand and analyse text data effectively. Figure 2 and Table 7 reflect the final changes.

Classification

After pre-processing, multiple classification models are implemented and then analysed which performs the best in identifying the positive and negative sentiments where:

- 1 is positive
- 0 is negative

Decision Tree

Figure 3

```
Decision Tree - Cross-Validation AUC_score: 0.739
Decision Tree - Cross-Validation Accuracy: 0.739
Decision Tree - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.79      0.75       500
           1       0.77      0.69      0.72       500

    accuracy                           0.74      1000
   macro avg       0.74      0.74      0.74      1000
weighted avg       0.74      0.74      0.74      1000

Decision Tree - Cross-Validation F1 Scores: [0.71379077 0.75494486 0.75494486 0.75361408 0.71898836]
Decision Tree - Cross-Validation Average F1 Score with variance: (0.739256585691811, 0.00035153814149218996)
```

Figure 4

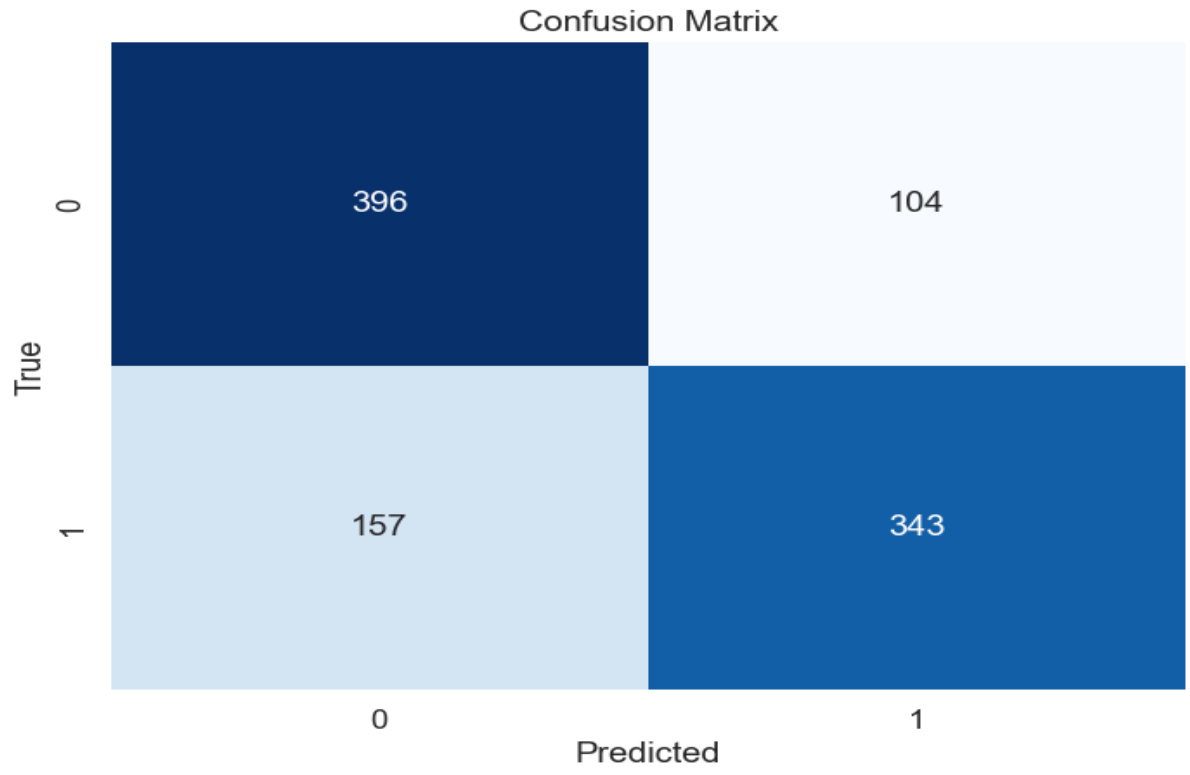
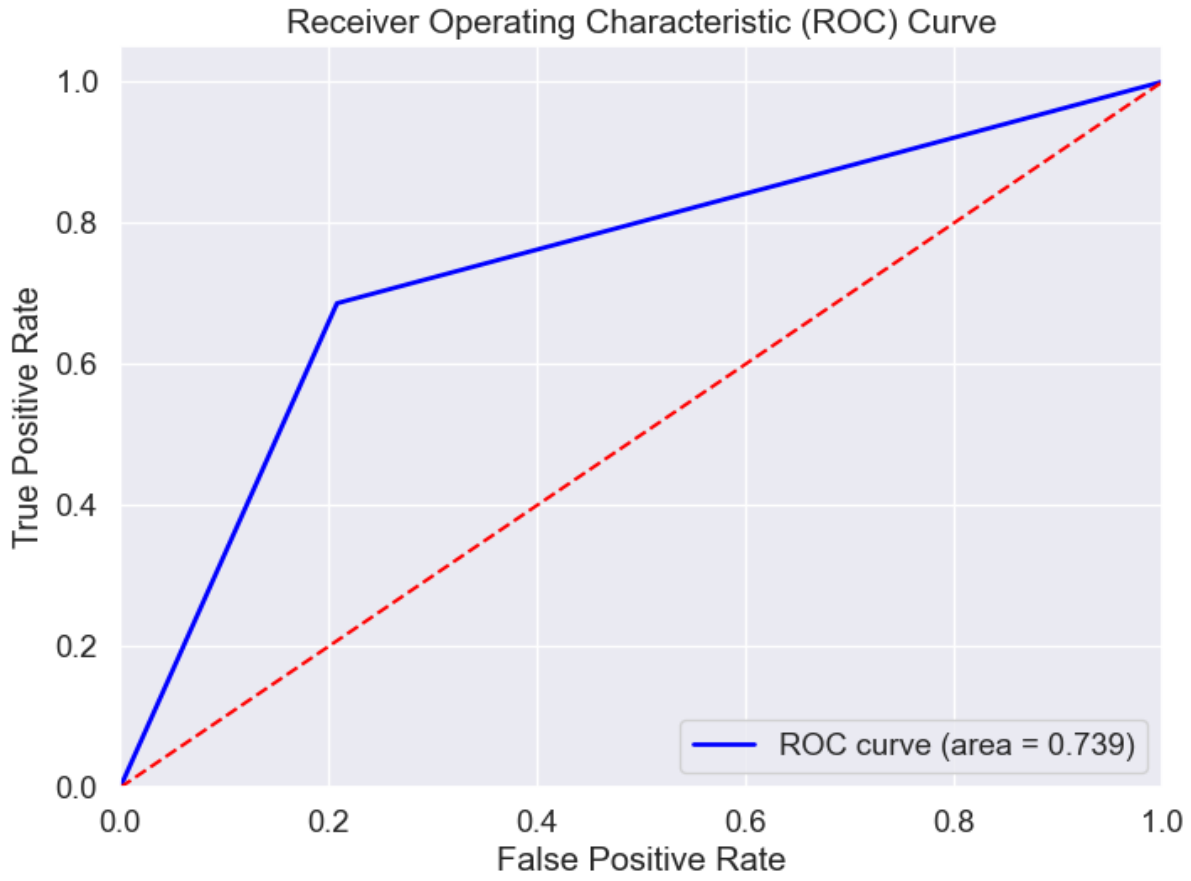


Figure 5



The Decision tree model achieves an F1 score and accuracy of around 0.73, which suggests a moderate ability to classify sentiments. The AUC score is also fair at 0.727. However, it tends to misclassify positive sentiments slightly more often than negative ones, as indicated by the confusion matrix.

Logistic Regression

Figure 6

```
Logistic - Cross-Validation AUC_score: 0.79
Logistic - Cross-Validation Accuracy: 0.79
Logistic - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.83      0.80       500
           1       0.82      0.75      0.78       500

    accuracy                           0.79      1000
   macro avg       0.79      0.79      0.79      1000
weighted avg       0.79      0.79      0.79      1000

Logistic - Cross-Validation F1 Scores: [0.85909823 0.81971154 0.77920514 0.75913288 0.82984686 0.76979281
 0.76886745 0.79991997 0.75757576 0.74937343]
Logistic - Cross-Validation Average F1 Score with variance: (0.789252406934822, 0.0011964002619988129)
```

Figure 7

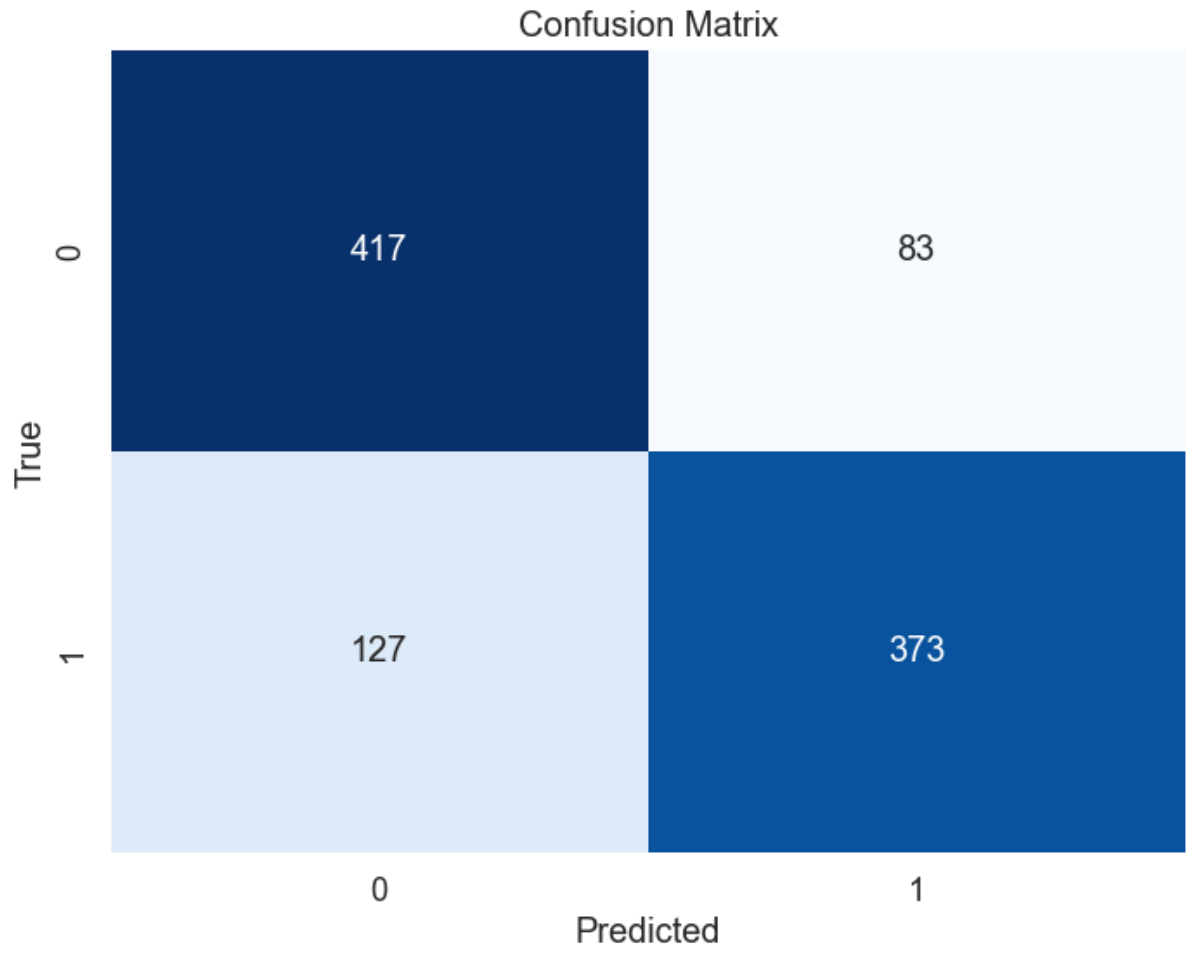
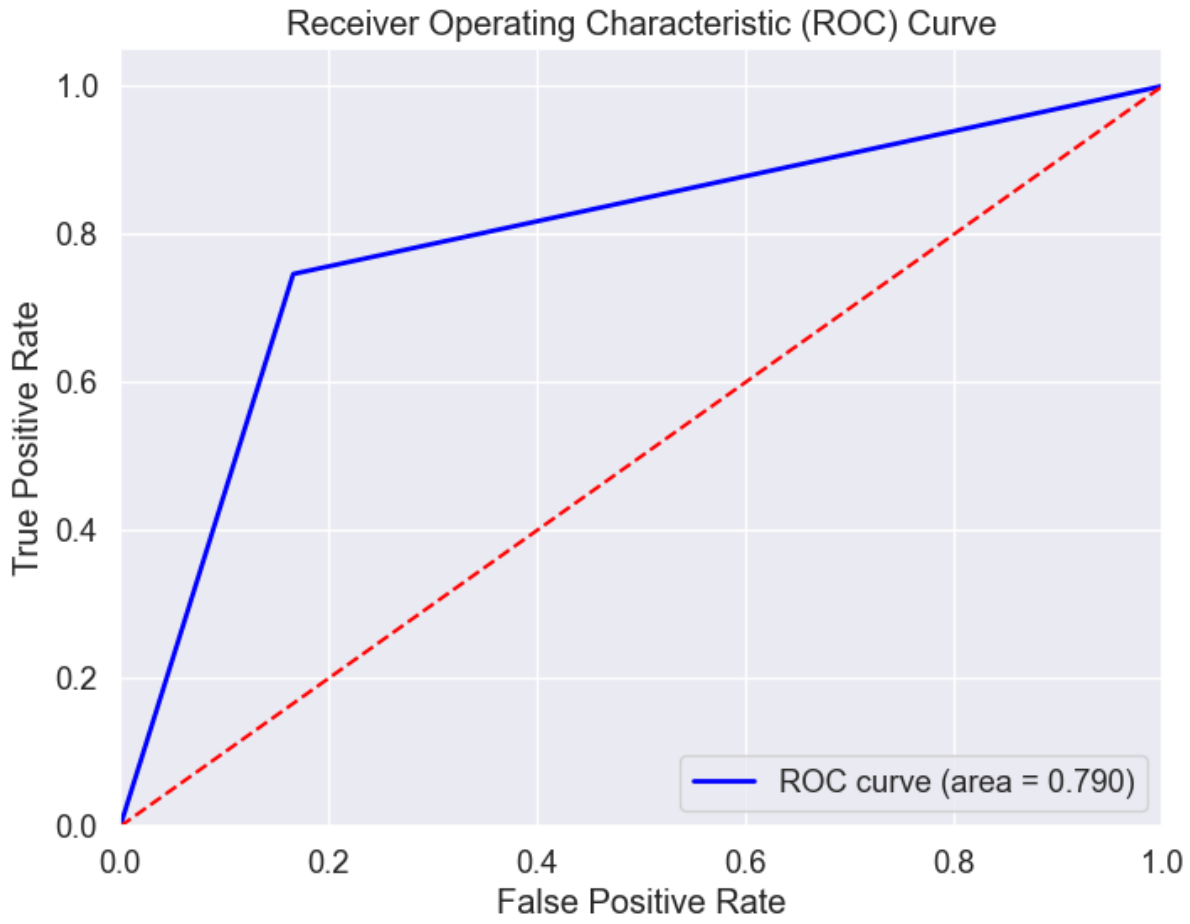


Figure 8



The Logistic Regression model performs slightly better than the Decision Tree, with an average F1 score and accuracy of around 0.79. It shows a balanced performance in classifying positive and negative sentiments, as shown by the confusion matrix. The AUC score is decent at 0.79.

SVM

Figure 9

```
SVM - Cross-Validation AUC_score: 0.7929999999999999
SVM - Cross-Validation Accuracy: 0.793
SVM - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.83      0.80       500
           1       0.82      0.75      0.78       500

    accuracy                           0.79      1000
   macro avg       0.79      0.79      0.79      1000
weighted avg       0.79      0.79      0.79      1000

SVM - Cross-Validation F1 Scores: [0.74937343 0.849985   0.78742788 0.74       0.74876897 0.83993597
 0.81884058 0.79967949 0.78       0.80952381]
SVM - Cross-Validation Average F1 Score with variance: (0.7923535128145446, 0.0013300193724349056)
```

Figure 10

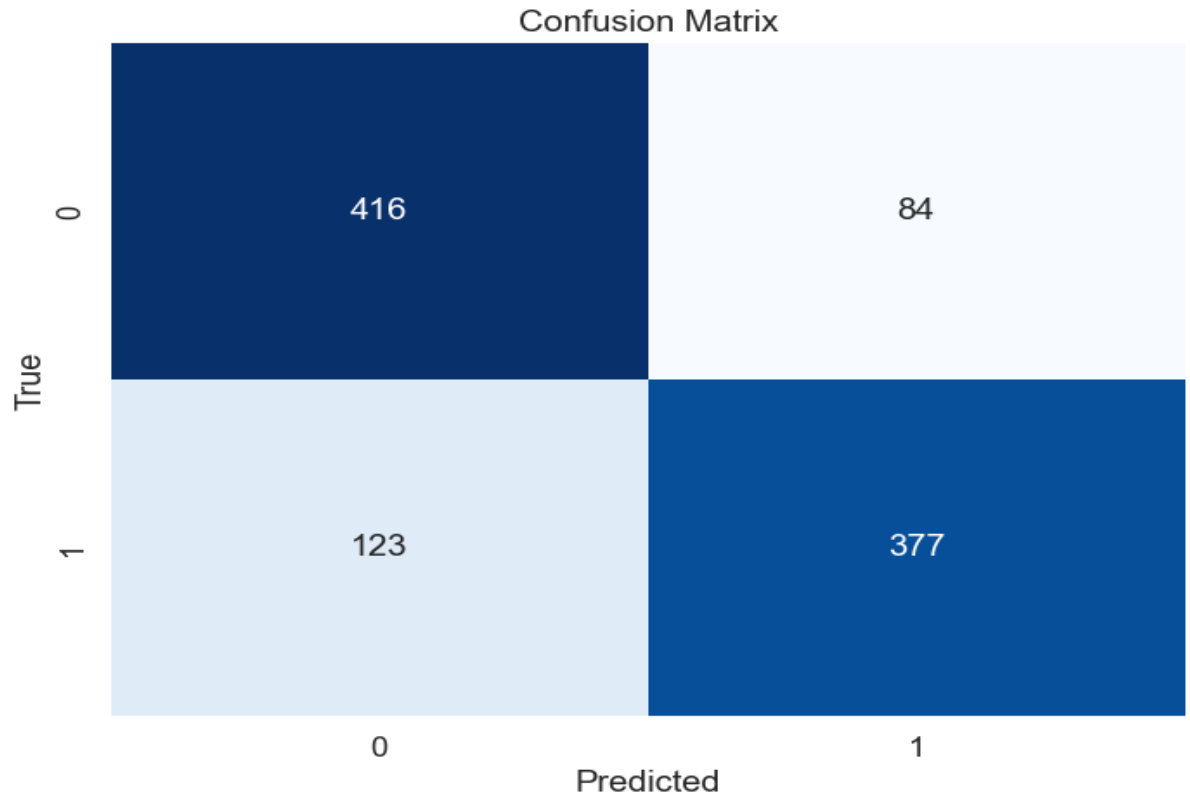
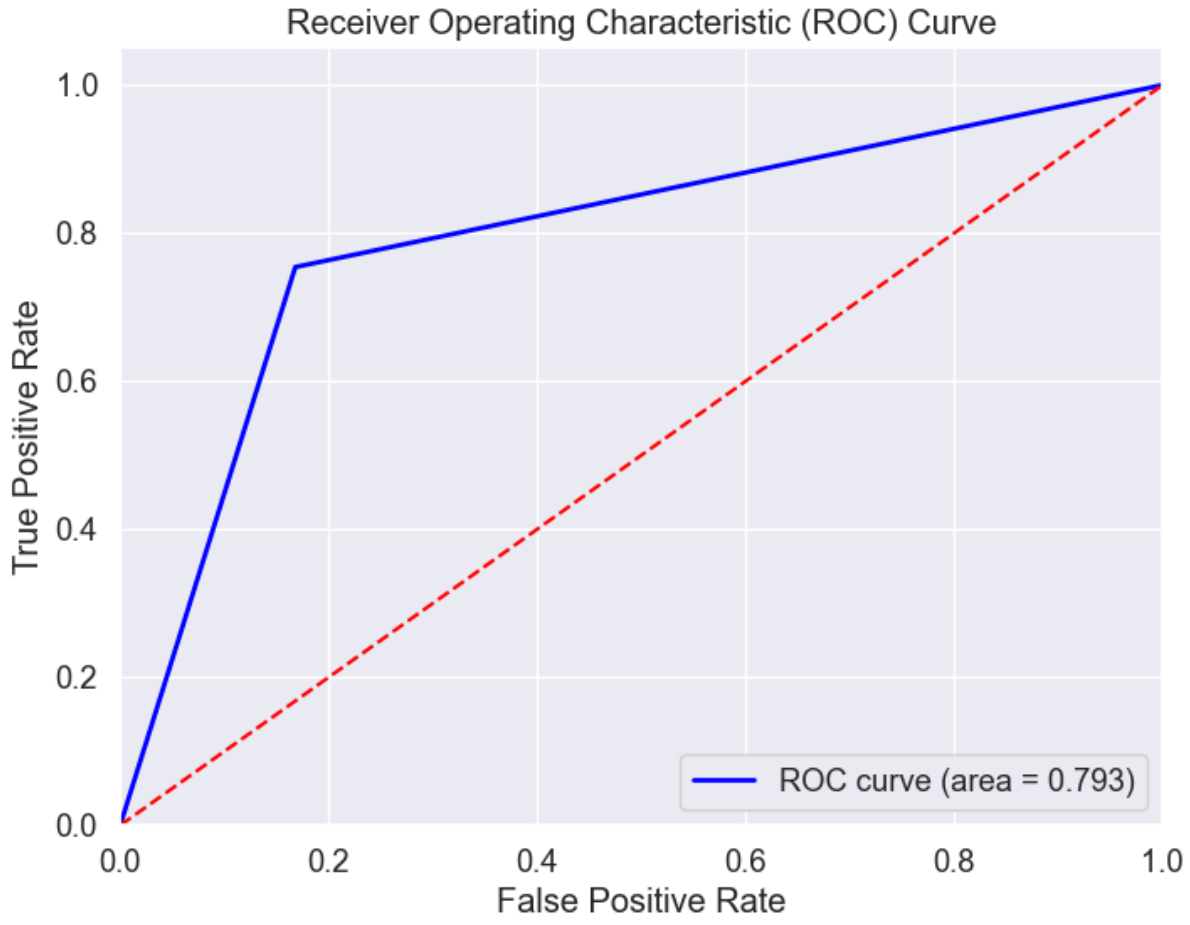

Confusion Matrix
416
84
123
377
0
1
True
0
1
Predicted

Figure 11


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.793)

Similar to Logistic Regression, SVM achieves an average F1 score and accuracy of around 0.79. It maintains a balanced prediction of positive and negative sentiments. The AUC score is also good at 0.793.

MLP

Figure 12

```
MLP - Cross-Validation AUC_score: 0.7630000000000001
MLP - Accuracy: 0.763
MLP - Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.77      0.77       500
           1       0.77      0.75      0.76       500

    accuracy                           0.76      1000
   macro avg       0.76      0.76      0.76      1000
weighted avg       0.76      0.76      0.76      1000

MLP - Cross-Validation F1 Scores: [0.81884058 0.769977   0.76       0.73958333 0.749975   0.75913288
 0.709971   0.78896593 0.729973   0.79967949]
MLP - Cross-Validation Average F1 Score with variance: (0.7626098201256082, 0.0009829744549610532)
```

Figure 13

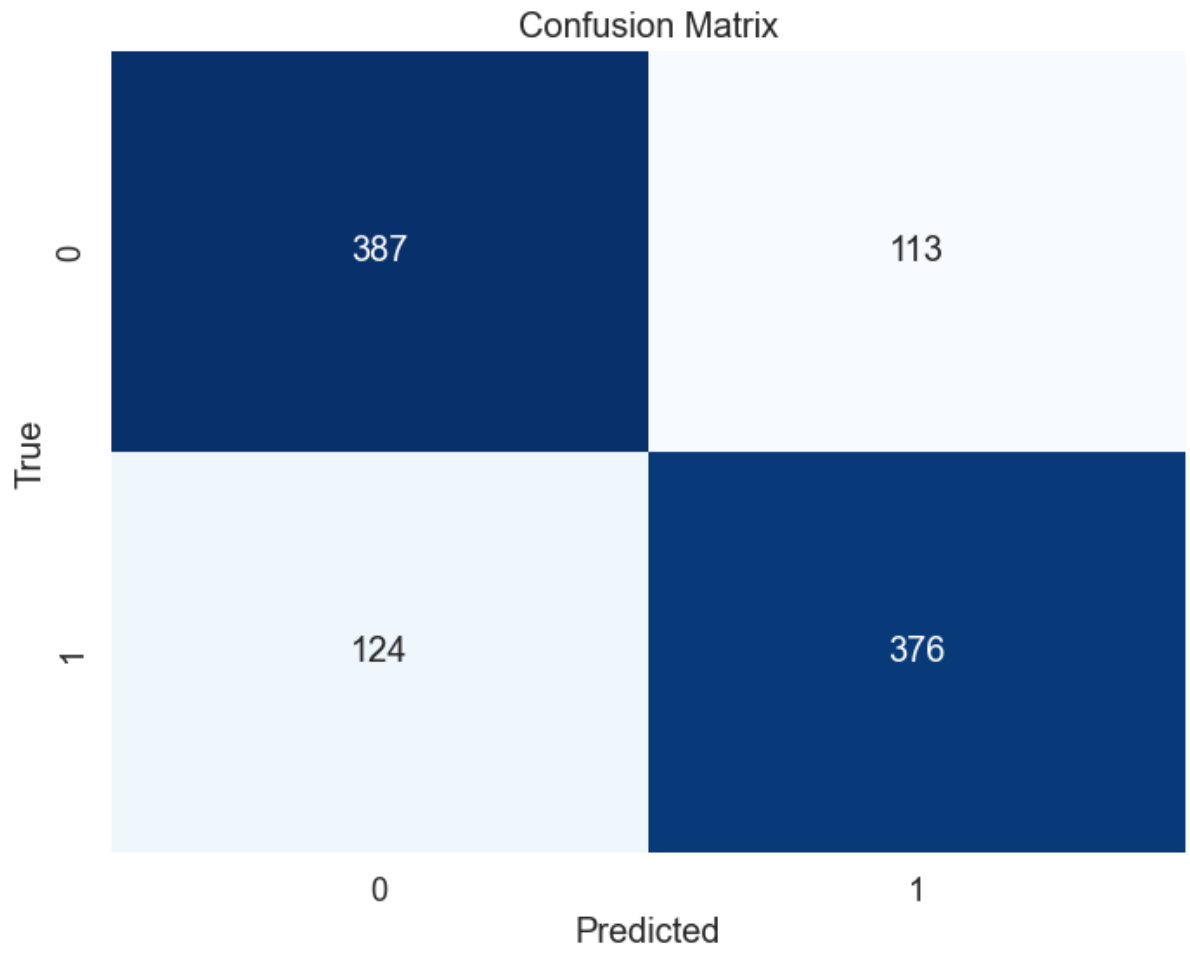
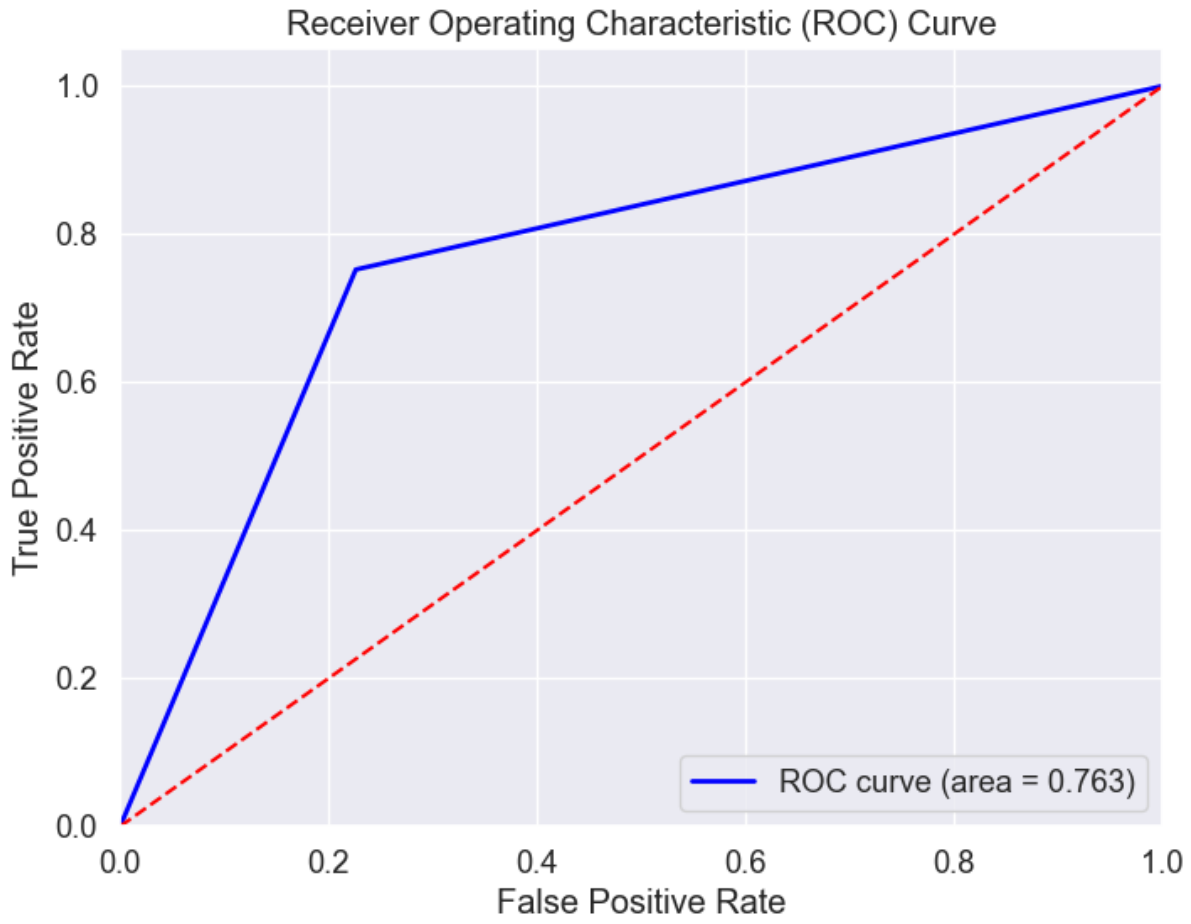


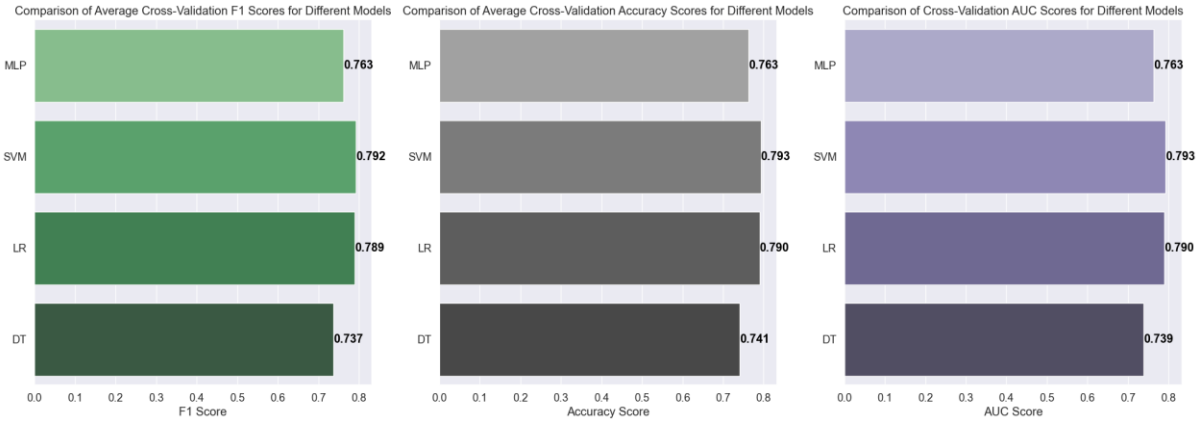
Figure 14



While achieving an F1 score and accuracy of around 0.76, MLP shows slightly lower performance compared to SVM and LR models. It struggles slightly more with precision and recall for both positive and negative sentiments. The AUC score is around 0.763, indicating moderate performance.

Comparison

Figure 15



The Decision Tree model shows a moderate performance with a slight bias in misclassifying positive sentiments. On the other hand, both Logistic Regression & SVM show similar and balanced

performances with higher accuracy and F1 scores compared to other models. Lastly, MLP shows a slightly lower performance in accurately classifying sentiments, especially when compared to Logistic Regression and SVM. In conclusion, Logistic Regression and SVM seem to be the better-performing models for sentiment classification in this scenario, displaying more balanced and higher predictive capabilities.

BERT-based classification

A large language model BERT is also implemented on the reviews dataset to classify positive and negative sentiments. Figure 16 reflects the “Loss” and “Accuracy” when training and testing the model. This essential means the model is trained the more accurate it gets as both training and validation trends are still positive with no over or under fitting.

Figure 16

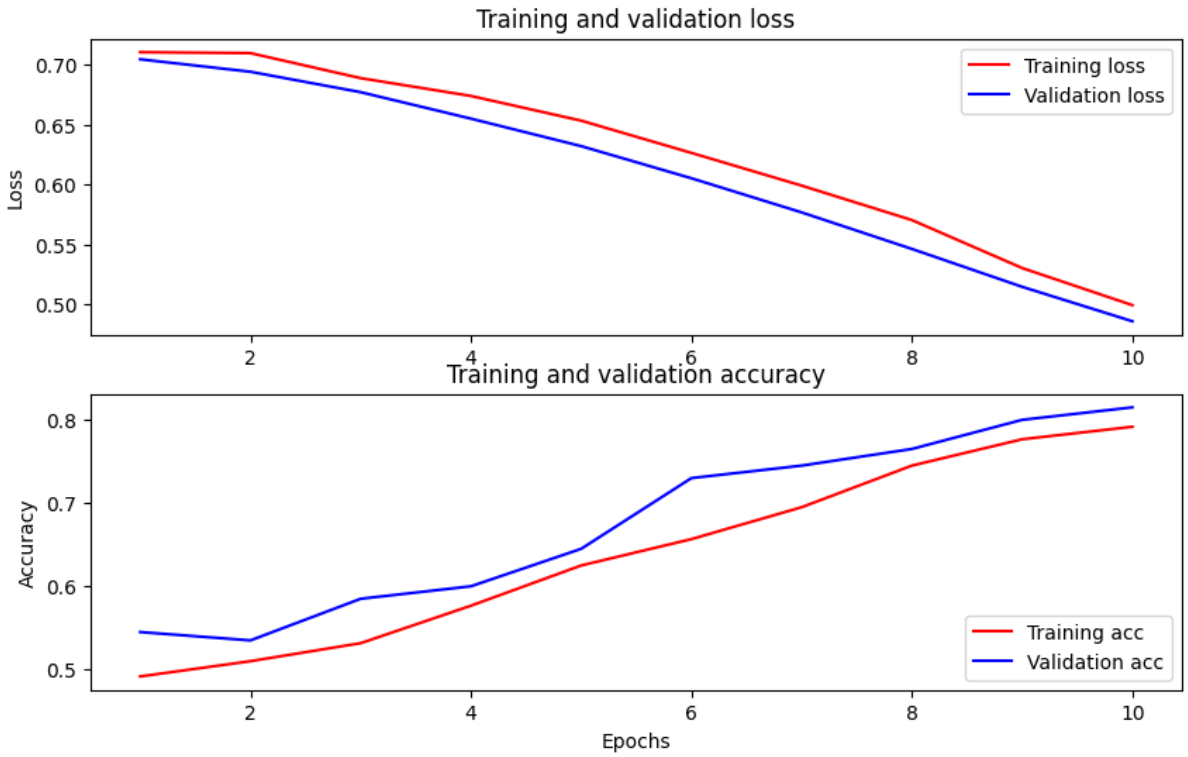


Figure 17

	precision	recall	f1-score	support
0	0.77	0.79	0.78	100
1	0.78	0.76	0.77	100
accuracy			0.78	200
macro avg	0.78	0.78	0.77	200
weighted avg	0.78	0.78	0.77	200

Accuracy: 0.775
AUC: 0.775
F1: 0.7715736040609137

Figure 18

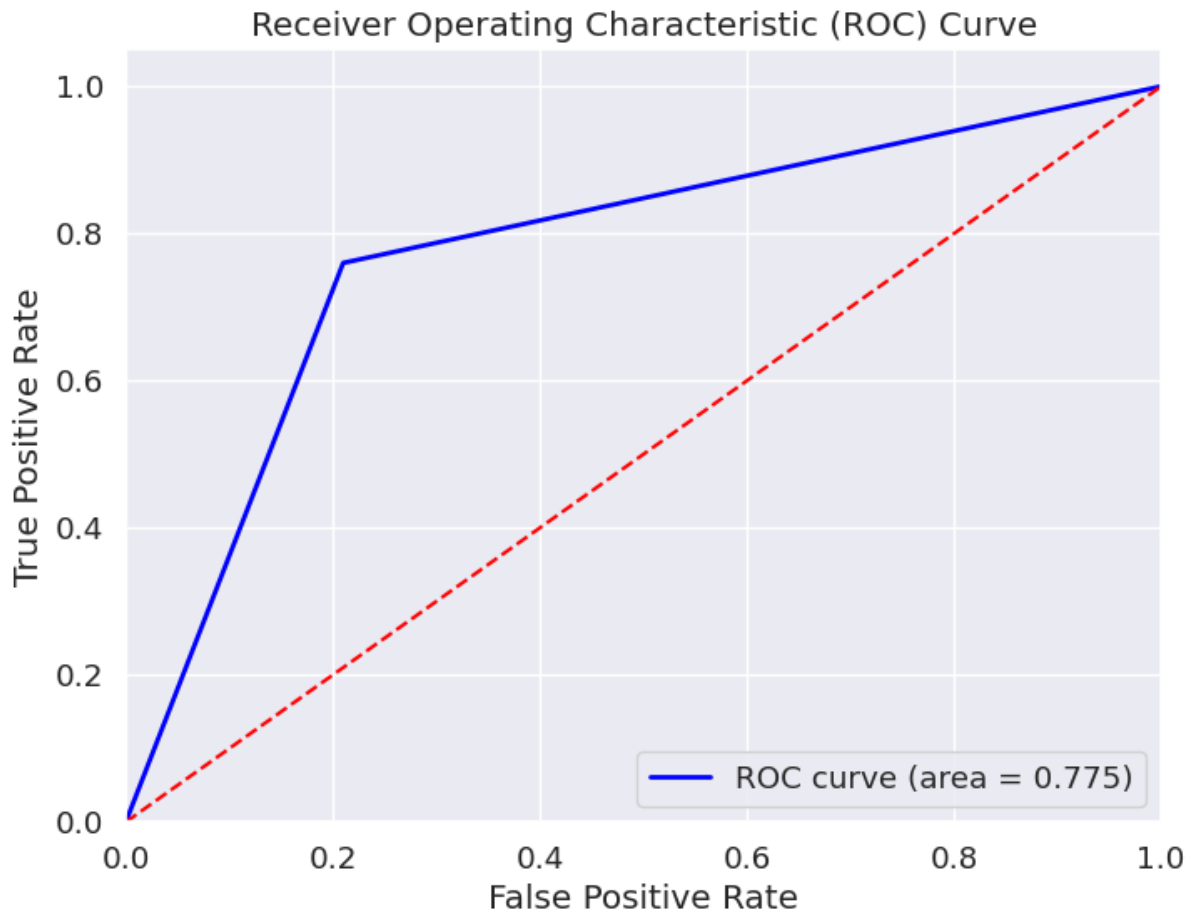
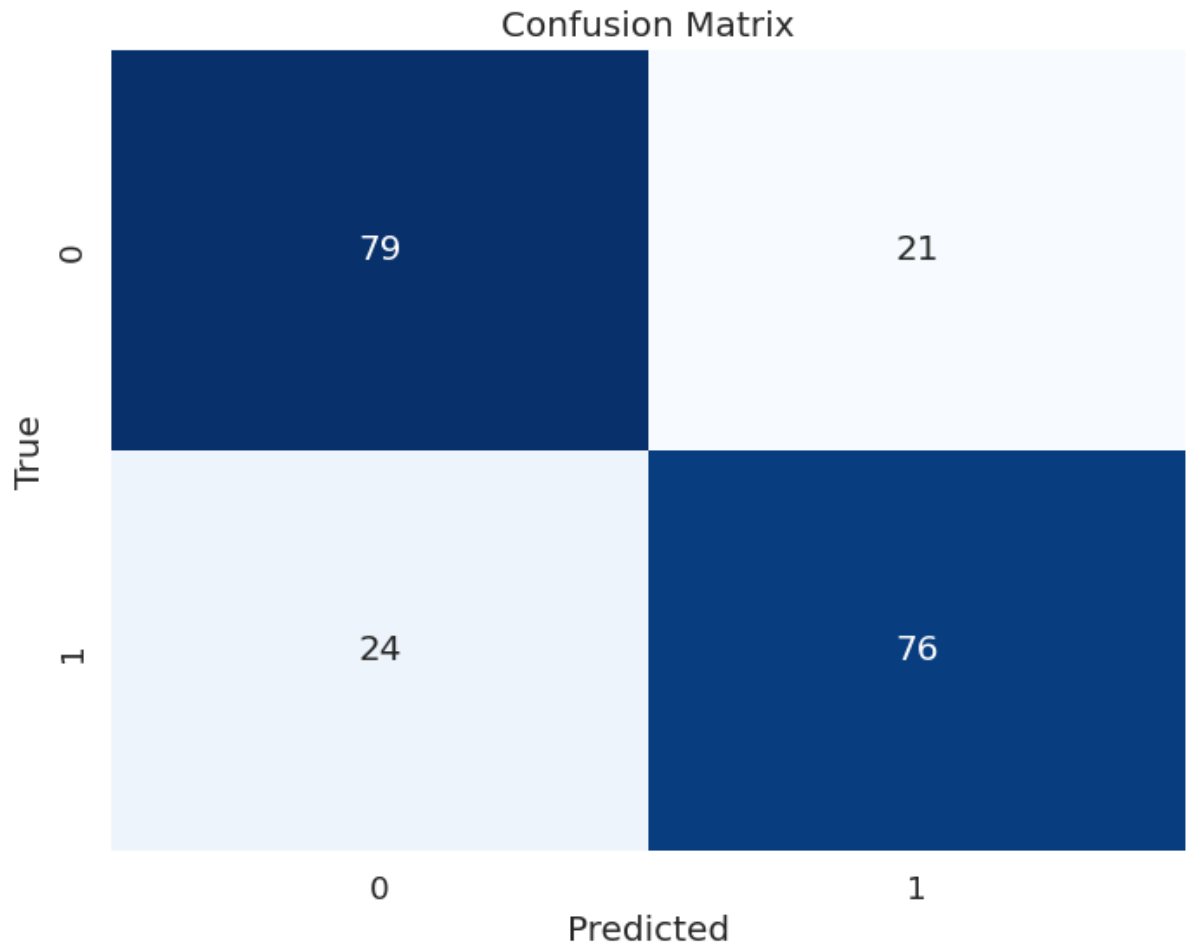


Figure 19



The BERT model achieves an accuracy and AUC score of 0.775, and an F1 score of approximately 0.772. It maintains a balanced precision and recall for both positive and negative sentiments, around 0.77 for each sentiment category. BERT demonstrates a fairly balanced prediction for positive and negative sentiments, with a slightly higher correct prediction rate for negative sentiments.

When compared to the other models BERT's performance metrics, particularly accuracy, F1 score, precision and recall align closely with Logistic Regression and SVM, which also scored around 0.79 in accuracy and F1 scores. This showcases a balanced prediction for positive and negative sentiments. In addition, BERT's AUC score and confusion matrix suggest a fairly balanced classification ability, similar to the other models discussed.

Overall, BERT's performance in sentiment analysis seems to align closely with the better-performing traditional models, such as Logistic Regression and SVM, showcasing comparable accuracy, precision, recall, and AUC scores. While BERT doesn't distinctly outperform these models in this specific scenario, it demonstrates similar capabilities in understanding sentiments within text data.

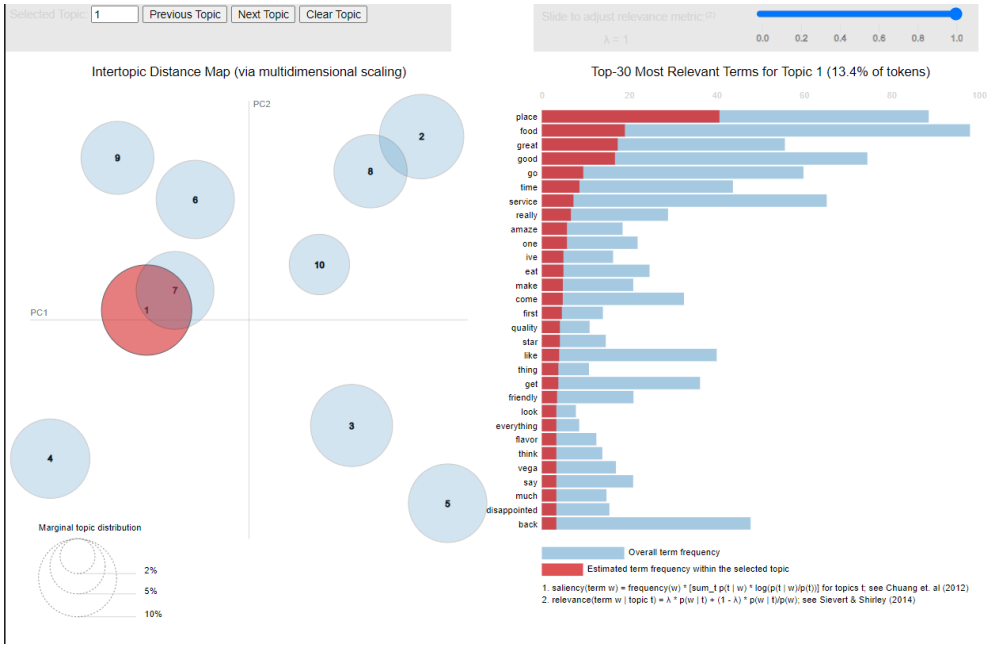
Topic Detection

LDA

Latent Dirichlet Allocation (LDA) is a tool that helps understand what people talk about in reviews or texts. It sorts words into different groups called "topics" based on how often they appear together. For instance, in a set of restaurant reviews, LDA might find topics like "good service and food" or "negative experiences." These topics give a quick idea of what most people liked or didn't like in their reviews, helping understand the main points without reading every single review.

Topic 1

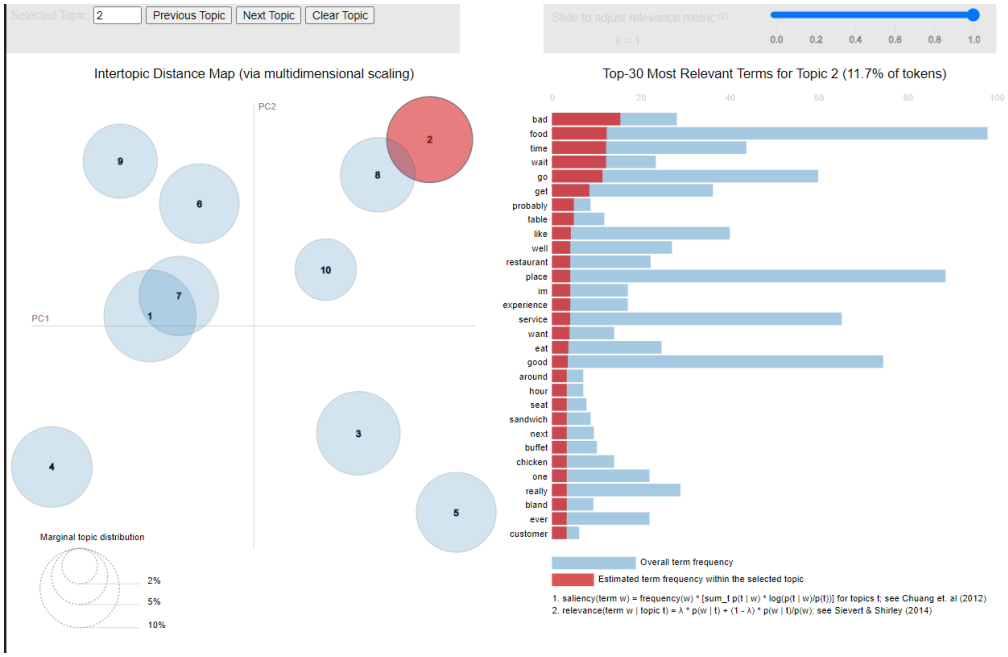
Figure 20



Topic 1 shows keywords include "service," "food," "good," "great," indicating positive sentiments about food and service quality.

Topic 2

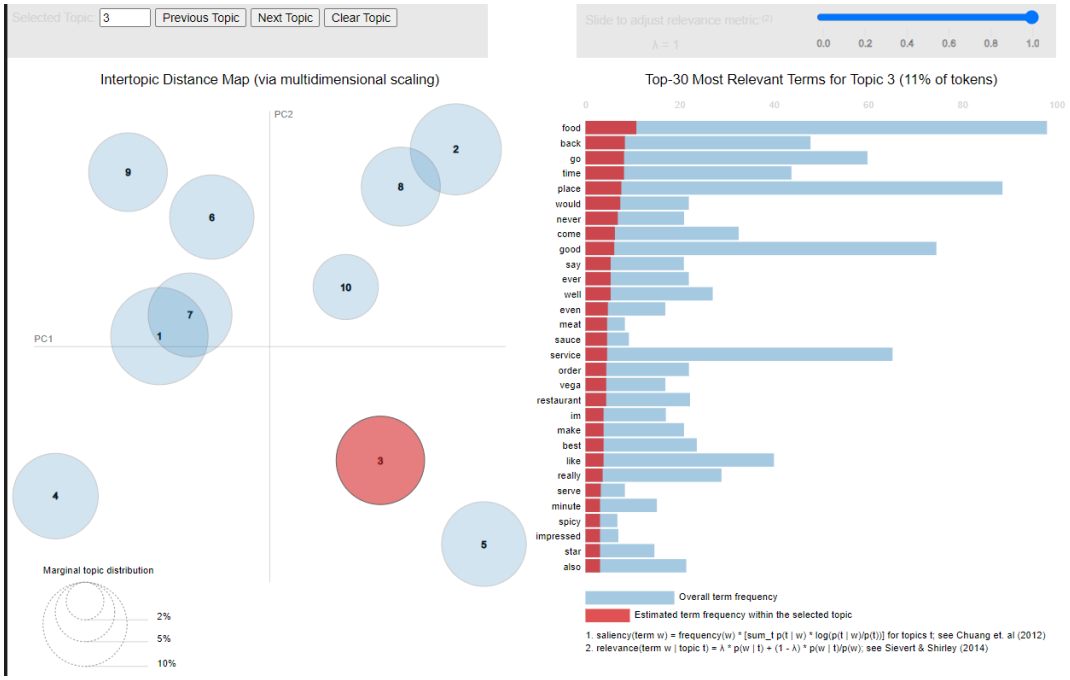
Figure 21



For this topic, words like "back," "never," "bad," suggest negative experiences and dissatisfaction with service or food quality.

Topic 3

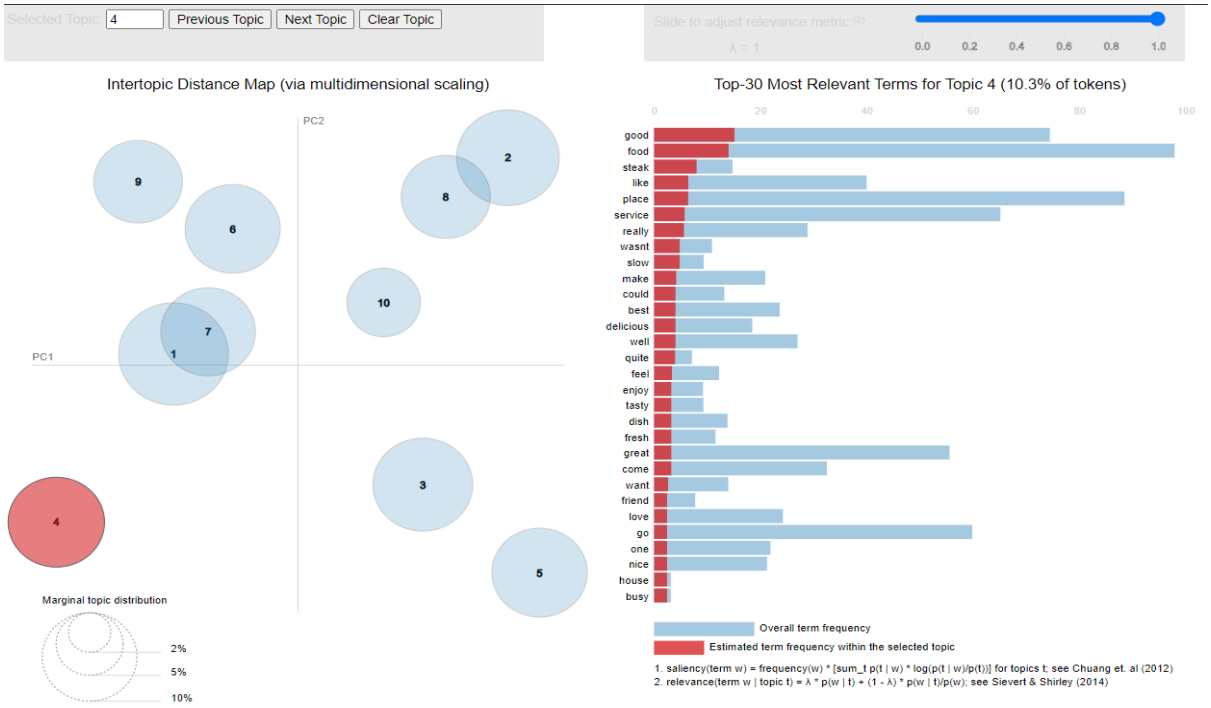
Figure 22



Positive sentiments with keywords like "good," "service," "great," indicating a positive dining experience.

Topic 4

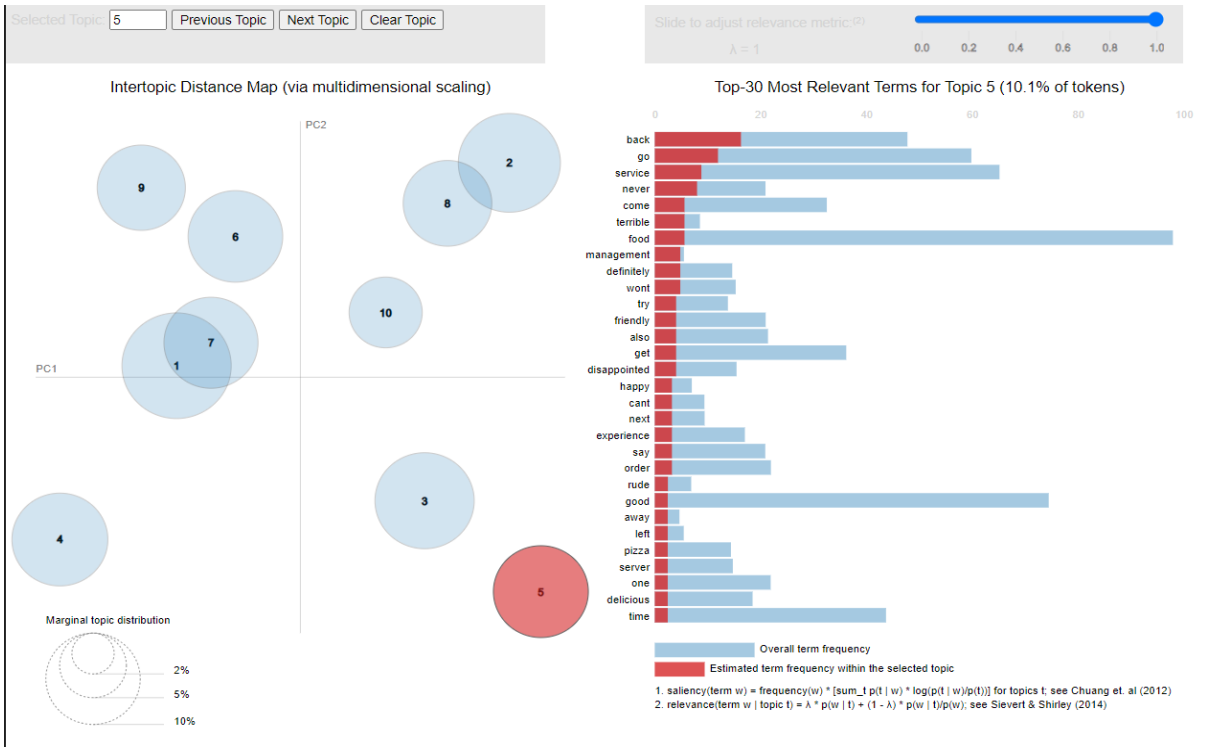
Figure 23



Keywords for this topic imply positive experiences regarding taste, freshness, and being the best place.

Topic 5

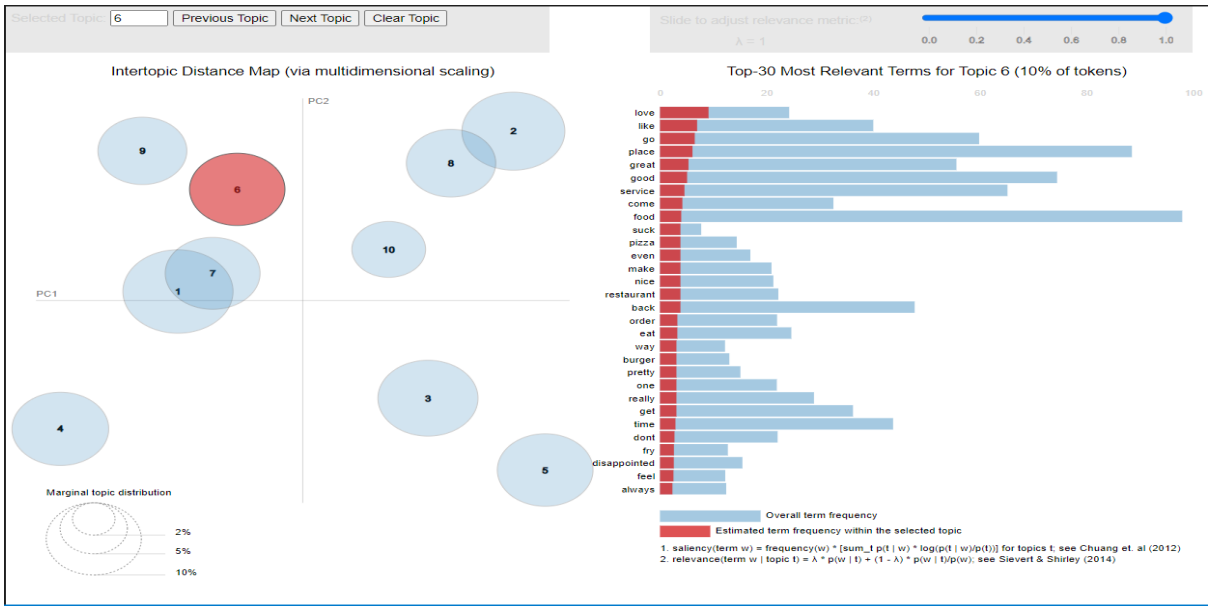
Figure 24



This topic Includes words like "bad," "wait," "terrible," indicating dissatisfaction and wait times.

Topic 6

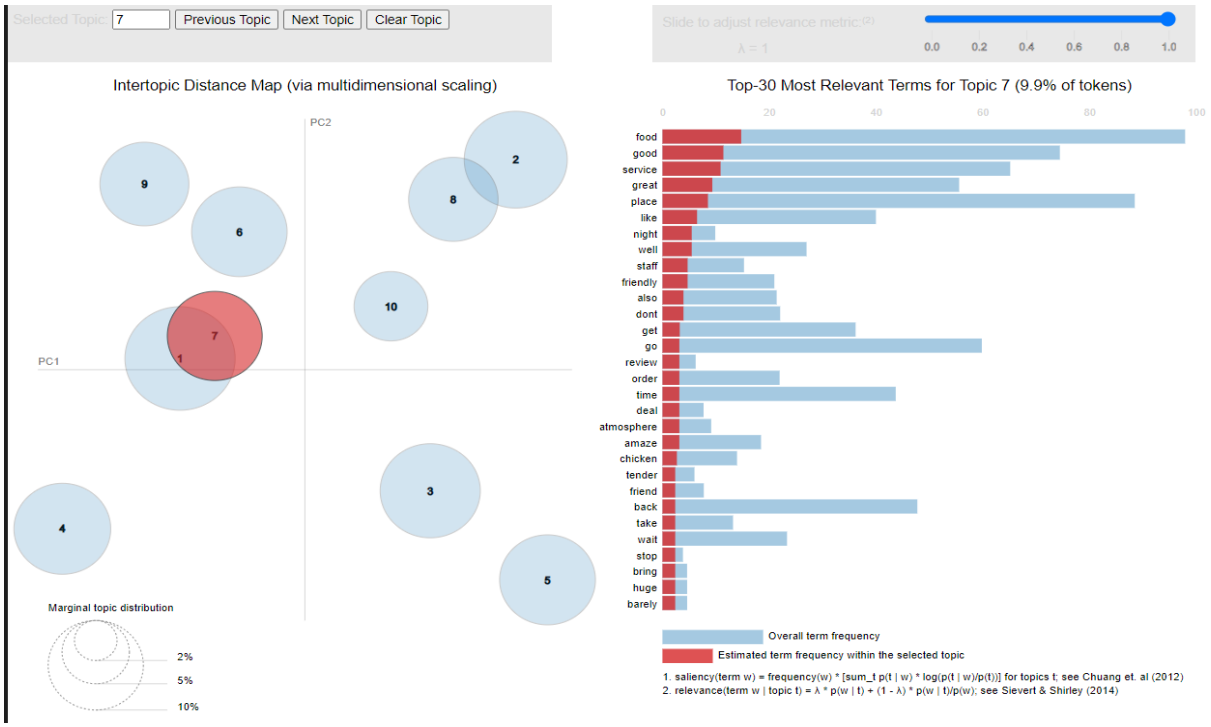
Figure 25



Words like "love," "great," "good," suggesting positive experiences and recommendations based on this Topic.

Topic 7

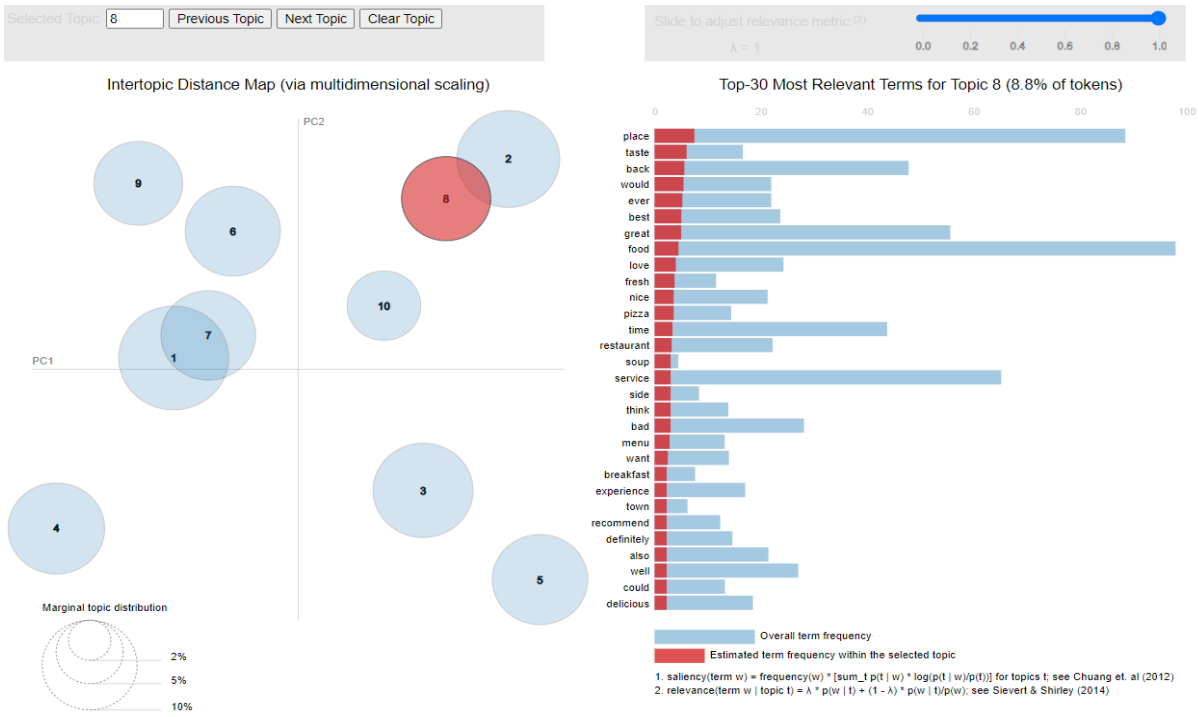
Figure 26



This topic focuses on food quality and mentions "steak" specifically.

Topic 8

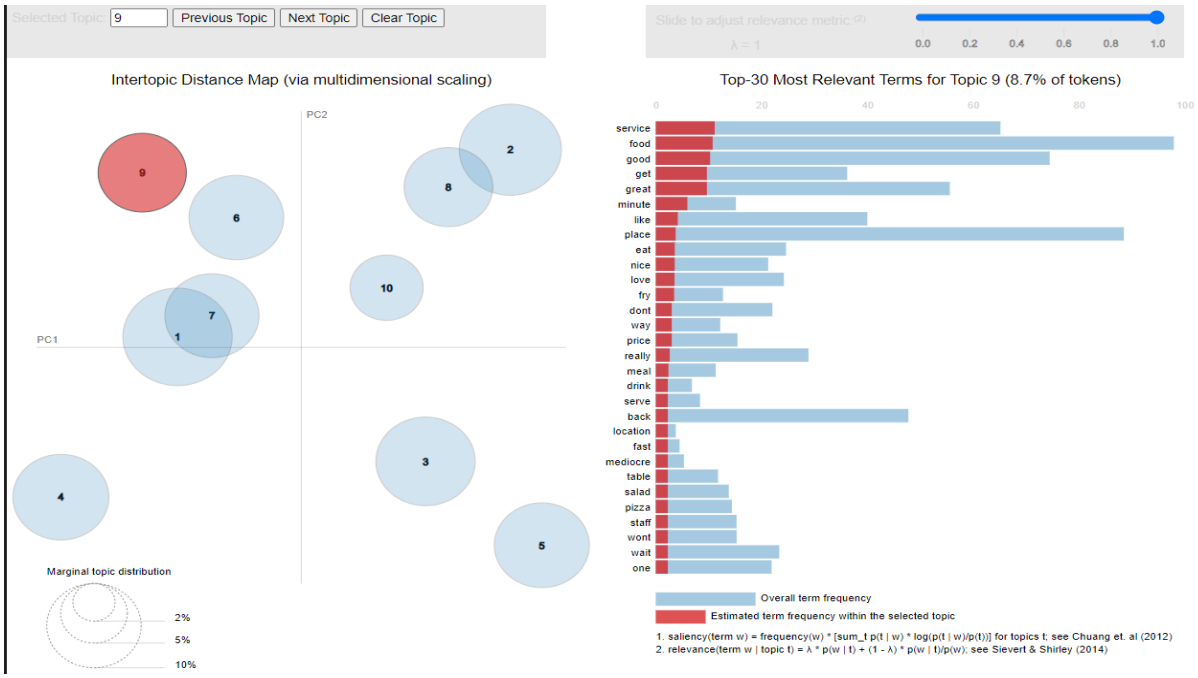
Figure 27



Topic 8 includes positive words like "good," "really," but also mentions slow service ("slow").

Topic 9

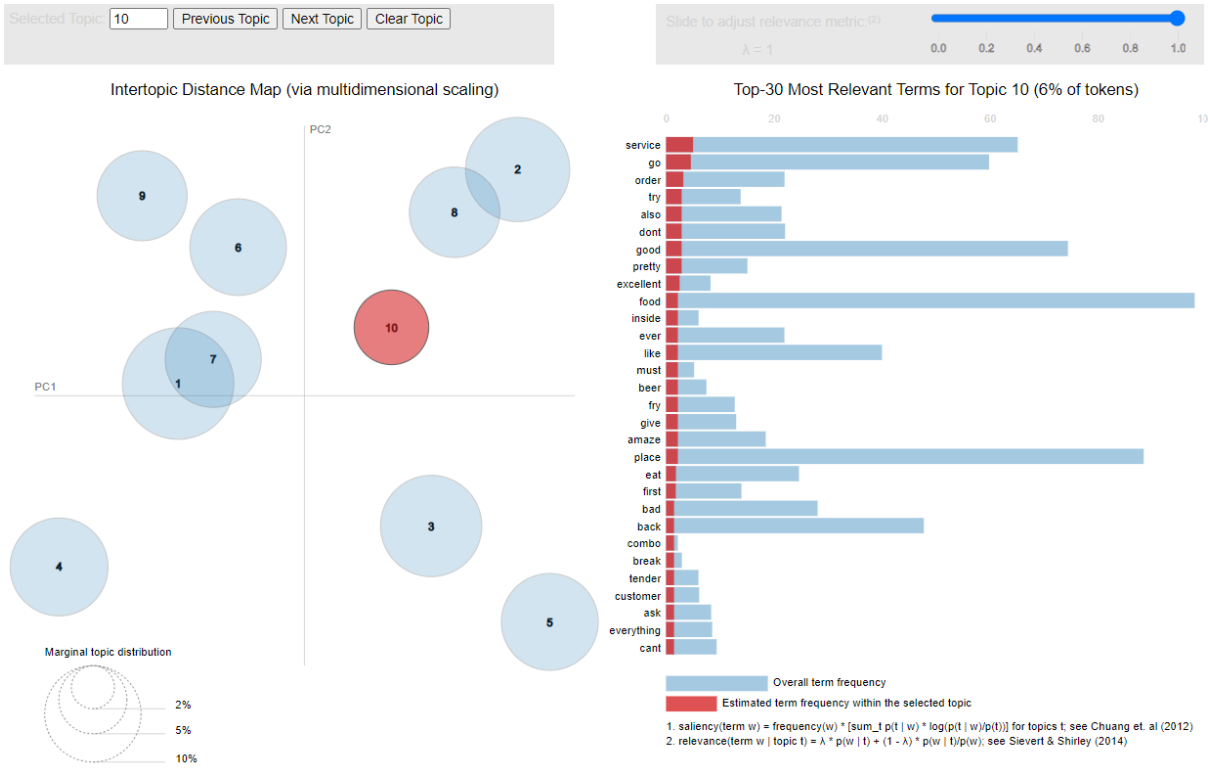
Figure 28



Keywords for topic 9 include "order," "service," and "try," possibly related to experiences with orders and service.

Topic 10

Figure 29



Lastly, topic 10 shows a positive sentiment about the place, food quality, and service.

LDA: Discussion and comparison

The topics cover various aspects such as service, food quality, waiting times, and overall experiences. This reflects a variety of topics related to food. Polarity variation is also shown as topics 1, 3, 4, 6, 8, and 10 convey positive sentiments, while topics 2, 5, 7, and 9 contain negative or mixed sentiments. In addition, there are several repetitive keywords. Certain words like "good," "great," "food," "service," and "place" appear across multiple topics, reflecting their significance in the reviews.

Lastly, the below shows the insights gathered from all 10 topics based on LDA:

- The analysis highlights both positive and negative aspects of the reviewed experiences, providing a comprehensive view of sentiments and topics expressed in the reviews.
- Some topics may represent recurring themes, such as positive sentiments about food quality and negative experiences related to waiting times or service.

LDA effectively helps in extracting meaningful topics from the reviews, allowing for a deeper understanding of the sentiments and themes expressed by customers.

NMF

Non-Negative Matrix Factorization (NMF) is a powerful mathematical technique employed in various fields for pattern recognition and feature extraction from multivariate data. Unlike LDA, NMF operates on non-negative values, making it particularly suited for tasks where data already exists in non-negative formats, such as in text mining. By factorizing an input matrix into two lower-dimensional matrices, NMF provides a unique way to decompose complex data into simpler, interpretable components.

In the context of restaurant reviews, NMF helps identify different topics customers talk about. It sorts feedback into groups like food quality, service, prices, and more. This helps managers understand what customers like or dislike, making it easier to improve specific areas like food or service for better customer satisfaction.

```
Topics and their top words:
Topic 1:
['good', 'price', 'selection', 'really', 'pizza']
Topic 2:
['place', 'love', 'recommend', 'like', 'eat']
Topic 3:
['food', 'delicious', 'bad', 'terrible', 'amaze']
Topic 4:
['service', 'friendly', 'slow', 'bad', 'fantastic']
Topic 5:
['great', 'time', 'experience', 'eat', 'staff']
```

In the NMF topic detection results for restaurant reviews, several distinct topics emerge based on prevalent words within each topic.

- Topic 1 seems to focus on aspects like value for money and food options, with words such as 'good', 'price', 'selection', and 'pizza'. It likely represents discussions about the quality and variety of food offerings in relation to their cost.

- Topic 2 captures sentiments and personal experiences, with words like 'place', 'love', 'recommend', and 'eat'. This topic appears to reflect customers' positive feelings and recommendations about the restaurant.

- Topic 3 reveals a polarity of opinions related to food quality, showcasing words like 'food', 'delicious', 'bad', 'terrible', and 'amaze'. This topic reflects varied opinions, showing both positive and negative sentiments about the taste and quality of the food.

- Topic 4 seems to be centred around customer service, featuring words like 'service', 'friendly', 'slow', 'bad', and 'fantastic'. It indicates diverse experiences with service quality, including positive attributes like friendliness alongside negative aspects like slowness.

- Topic 5 revolves around the overall dining experience, highlighting words such as 'great', 'time', 'experience', 'eat', and 'staff'. This topic seems to encapsulate customers' holistic experiences, including their time at the restaurant, the dining experience, and interactions with the staff.

Each topic captures different facets of the dining experience ranging from food quality, service, and personal sentiments. This provides a better view of various aspects highlighted in the customers' reviews.

Comparing the results of LDA to NMF

The analysis using LDA, highlighting various topics. It covers aspects like service, food quality, waiting times, and overall experiences. It emphasises polarity variations across topics, indicating positive sentiments in some and negative or mixed sentiments in others. Additionally, it notes repetitive keywords across different topics and emphasises the effectiveness of LDA in extracting meaningful themes.

On the other hand, NMF's results show distinct topics in restaurant reviews. It outlines specific themes captured in each topic, such as value for money and food options, sentiments and personal experiences, opinions on food quality, customer service, and the overall dining experience. This description focuses on the diversity of aspects covered by each topic, providing a detailed view of various facets highlighted in customer reviews.

In conclusion, both LDA and NMF methods give different but useful insights. LDA shows a bigger picture of how people feel about things like service and food quality. NMF looks closer at specific things within these topics, giving more details. Together, they help restaurants know what customers like and don't like, so they can make things better and keep customers happy. Mixing these methods helps see the whole story behind what customers say. It is also important to note that the number of topics chosen for LDA is 10 and NMF is 5. This difference affects how detailed the analysis gets and what specific aspects it focuses on.