TABLE OF CONTENTS

1. PREPROCESSING TEXTUAL DATA

1.1. OVERVIEW

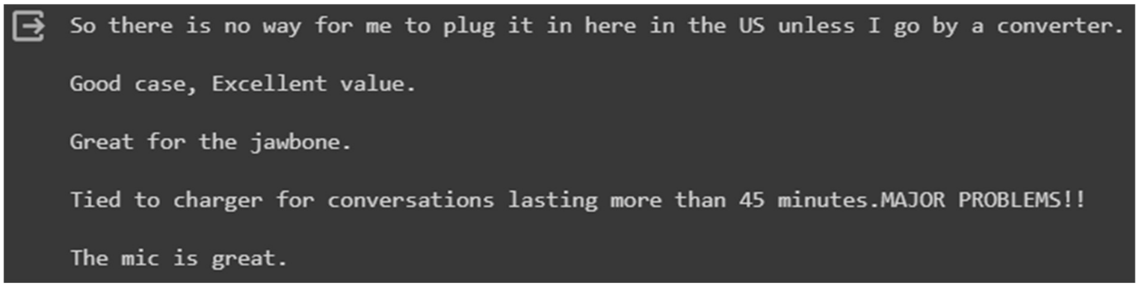
Effective text preprocessing is essential in the field of natural language processing (NLP) to derive valuable insights from unprocessed textual data. Task 1 of this course focuses on the careful preparation of Amazon review data using a variety of preprocessing methods to guarantee that the data used for the ensuing analyses is accurate and meaningful.

1.2. METHODOLOGY

1.2.1. REMOVING PUNCTUATION

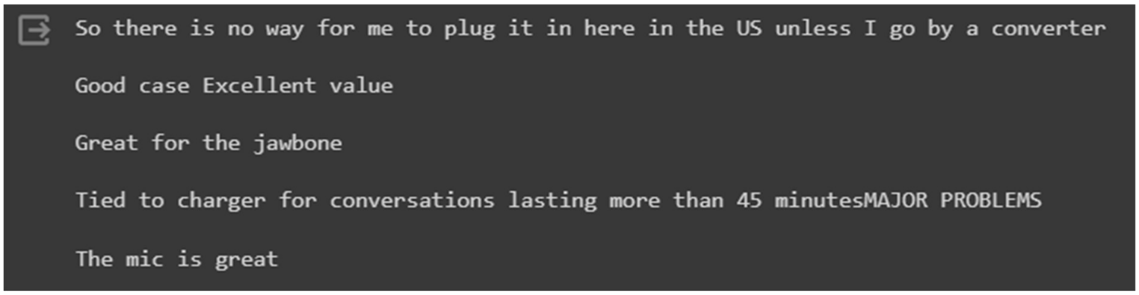
The removal of punctuation marks facilitates sentence streamlining by getting rid of extraneous symbols that don't significantly improve sentiment analysis or subject identification.

SENTENCES BEFORE REMOVING PUNCTUATION



```
So there is no way for me to plug it in here in the US unless I go by a converter.

Good case, Excellent value.

Great for the jawbone.

Tied to charger for conversations lasting more than 45 minutes.MAJOR PROBLEMS!!

The mic is great.
```

SENTENCES AFTER REMOVING PUNCTUATION



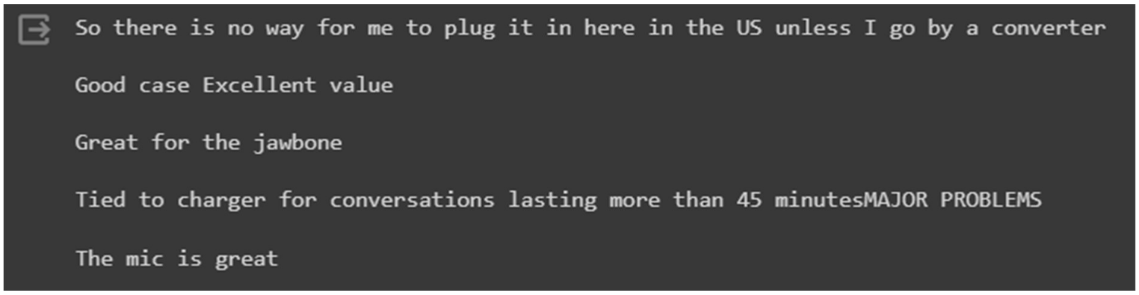
```
So there is no way for me to plug it in here in the US unless I go by a converter

Good case Excellent value

Great for the jawbone

Tied to charger for conversations lasting more than 45 minutesMAJOR PROBLEMS

The mic is great
```

It is evident from the above figures that each of the five statements has had the punctuation marks (,.!) removed from it.

1.2.2. REMOVING NUMBERS

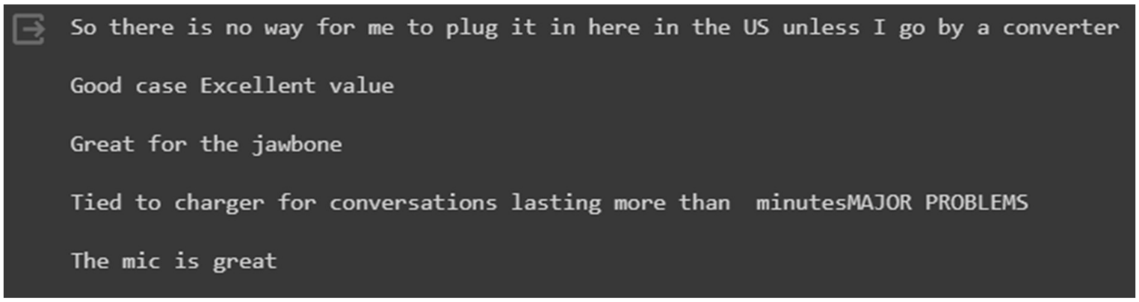
To minimise noise and enable a more precise examination of the text, numerical digits that were frequently unconnected to sentiment interpretation were methodically eliminated.

SENTENCES BEFORE REMOVING NUMBERS



```
So there is no way for me to plug it in here in the US unless I go by a converter

Good case Excellent value

Great for the jawbone

Tied to charger for conversations lasting more than 45 minutesMAJOR PROBLEMS

The mic is great
```

SENTENCES AFTER REMOVING NUMBERS



```
So there is no way for me to plug it in here in the US unless I go by a converter

Good case Excellent value

Great for the jawbone

Tied to charger for conversations lasting more than  minutesMAJOR PROBLEMS

The mic is great
```

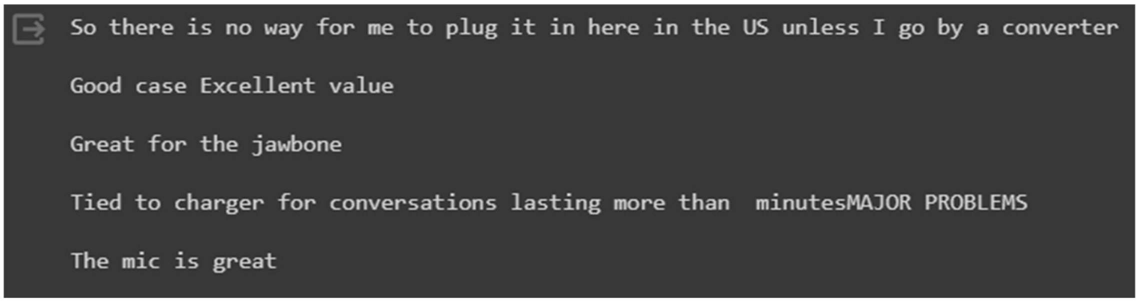
The function was applied, and as can be seen in the photos above, the fourth sentence—which had the number 45—was eliminated.

1.2.3. REMOVING STOP WORDS

The phrases were edited to remove common English stop words. This is an important step in helping the algorithm better identify sentiment subtleties by concentrating on words that carry content.

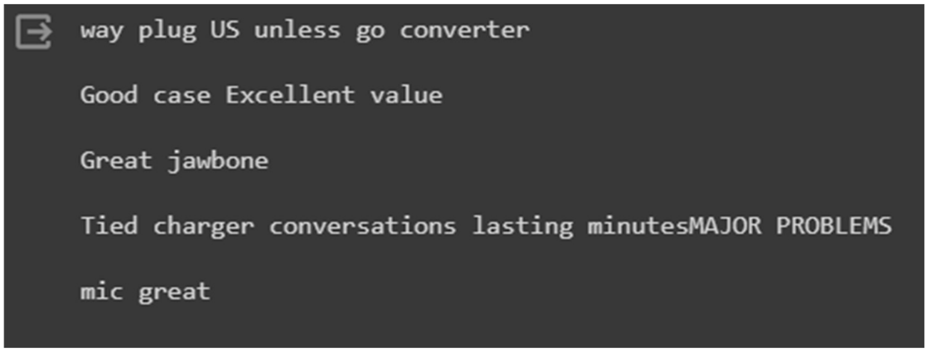
The first sentence has been cleared of stop words like 'the', 'a', 'by' 'is', 'so', 'no', 'i', 'in', and 'for'. No words are eliminated from the second sentence because it contains no stop words. The third sentence has words like 'for' and 'the' removed. Stop words like 'to', 'for', 'more', and 'than' are absent from the fourth sentence, and the words 'the' and 'is' are eliminated from the final sentence.

SENTENCES BEFORE REMOVING STOP WORDS



```
So there is no way for me to plug it in here in the US unless I go by a converter

Good case Excellent value

Great for the jawbone

Tied to charger for conversations lasting more than  minutesMAJOR PROBLEMS

The mic is great
```

SENTENCES AFTER REMOVING STOP WORDS

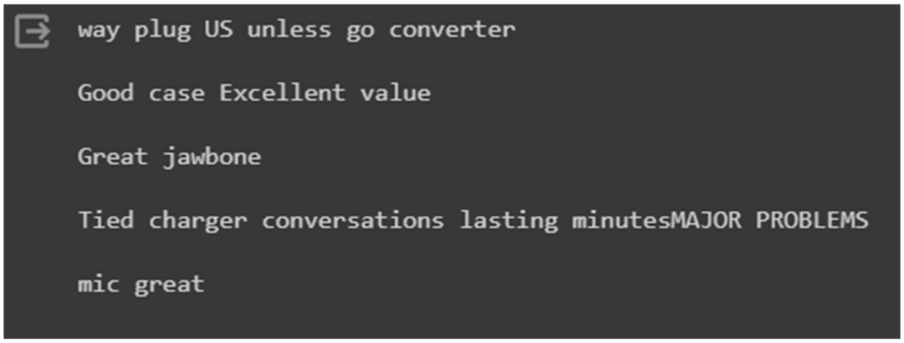


```
way plug US unless go converter

Good case Excellent value

Great jawbone

Tied charger conversations lasting minutesMAJOR PROBLEMS

mic great
```

1.2.4. CHANGING TEXT TO LOWER CASE

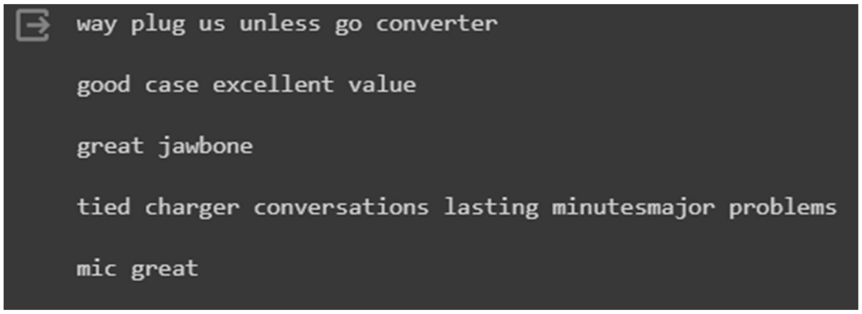
By making the text lowercase, the model is prevented from recognising words with different cases as separate entities, ensuring uniformity and consistency. As we can see from the below figures, the words have all their letters changed to lower case.

SENTENCES BEFORE APPLYING LOWER CASE



```
way plug US unless go converter

Good case Excellent value

Great jawbone

Tied charger conversations lasting minutesMAJOR PROBLEMS

mic great
```

SENTENCES AFTER APPLYING LOWER CASE



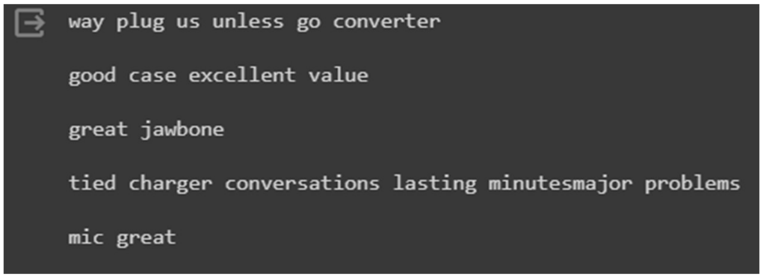
```
way plug us unless go converter

good case excellent value

great jawbone

tied charger conversations lasting minutesmajor problems

mic great
```

1.2.5. LEMMATIZING WITH PART-OF-SPEECH (POS) TAGGING

To capture the essence of a word and facilitate the discovery of sentiment or topic-related patterns, lemmatization entails breaking down words into their base or root forms. Furthermore, each word is assigned a grammatical category by use of POS tagging, which guarantees precise lemmatization according to the function of the word in the sentence.

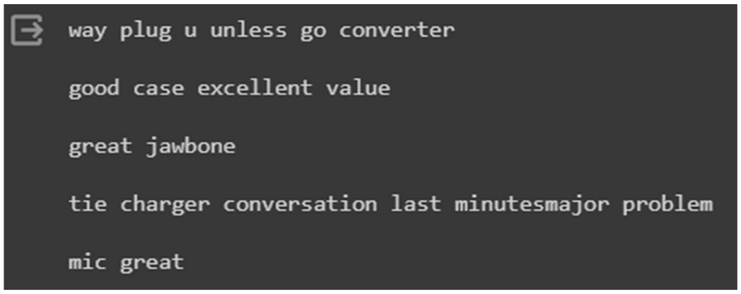
The following picture makes it evident how words like "tied" are converted to their base form "tie," just as "lasting" is changed to "last," "problems" to "problem," and "conversations" to "conversation."

SENTENCES BEFORE LEMMATIZING



```
way plug us unless go converter

good case excellent value

great jawbone

tied charger conversations lasting minutesmajor problems

mic great
```

SENTENCES AFTER LEMMATIZING

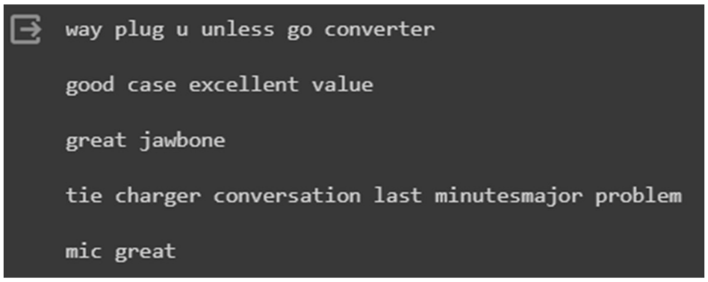


```
way plug u unless go converter

good case excellent value

great jawbone

tie charger conversation last minutesmajor problem

mic great
```

1.2.6. REMOVING SHORT WORDS

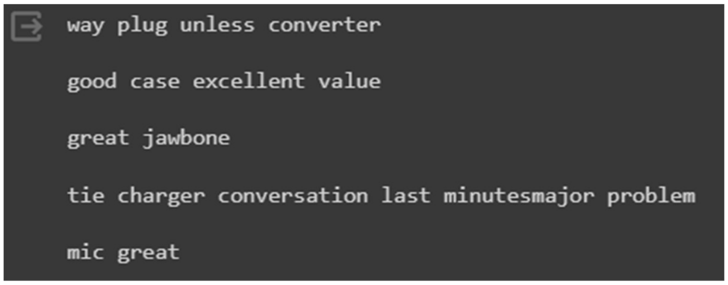
The first sentence, as seen in the visual below, contains the letter "u," which, before any preprocessing techniques are used, refers to the country US. However, the word was reduced to a single letter that had nothing to do with the original word after several procedures were used. Thus, a strategy for eliminating short terms with two letters or less is used.

SENTENCES BEFORE REMOVING SHORT WORDS



```
way plug u unless go converter

good case excellent value

great jawbone

tie charger conversation last minutesmajor problem

mic great
```

SENTENCES AFTER REMOVING SHORT WORDS



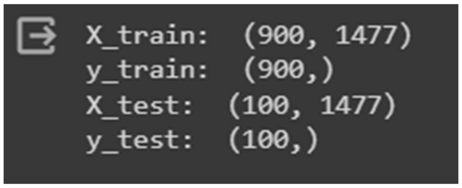
```
way plug unless converter

good case excellent value

great jawbone

tie charger conversation last minutesmajor problem

mic great
```

2. CLASSIFICATION USING BAG-OF-WORDS

2.1. OVERVIEW

In this study, a bag-of-words representation was used to perform sentiment analysis on Amazon evaluations. The main goal was to identify the tone of customer reviews and classify them as favourable or negative. To elucidate specific insights and evaluate the effectiveness of each model, four classification algorithms were used: Multinomial Naive Bayes (MNB), Support Vector Classifier (SVC), Random Forest Classifier (RFC), and an Artificial Neural Network (ANN).

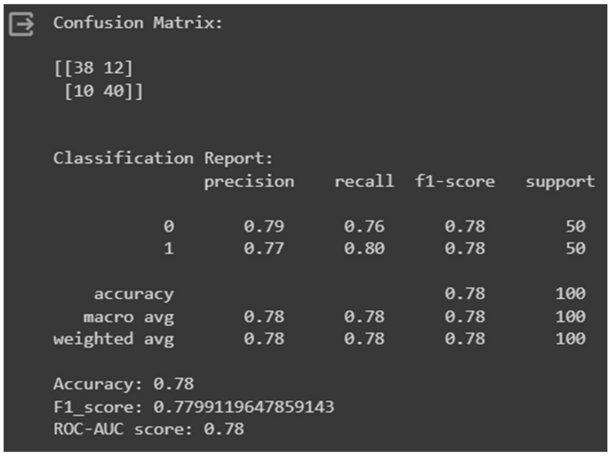
The dataset, which was derived from reviews on Amazon, included labelled sentences that represented feelings, with '0' denoting negativity and '1' positive sentiment. The dataset is split into train and test set, with test ratio of 0.1 (10%). Each set's shapes are shown below.



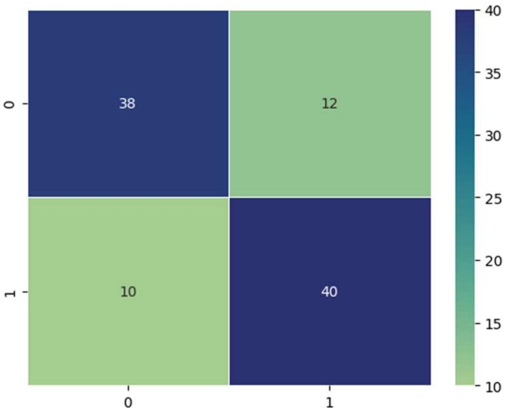
```
X_train:  (900, 1477)
y_train:  (900,)
X_test:  (100, 1477)
y_test:  (100,)
```

2.2. MULTINOMIAL NAÏVE BAYES

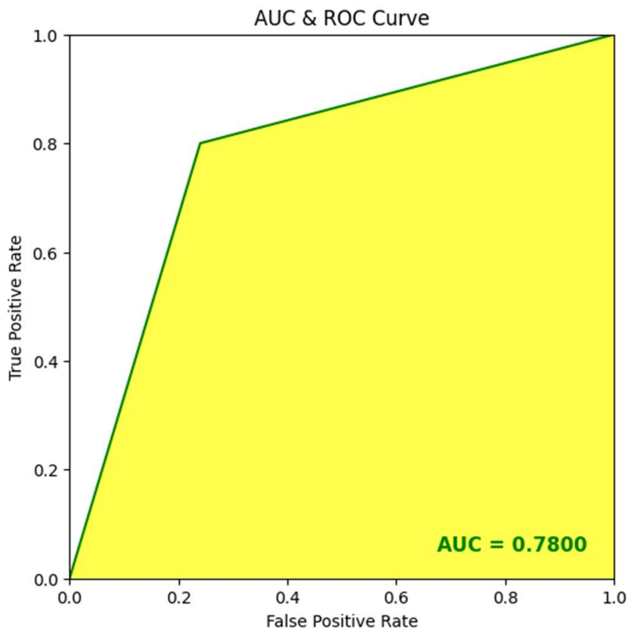
By using MNB, we were able to obtain an impressive accuracy of 78%. To fine-tune hyperparameters, a grid search with cross-validation was carried out, producing the ideal values {'alpha': 1.0, 'class_prior': None, 'fit_prior': True}. A balanced classification performance was found in the confusion matrix, with 38 true negatives, 40 true positives, 12 false positives, and 10 false negatives. The model's efficacy in identifying sentiments was further supported by the precision and recall metrics as shown in below figures.



```
Confusion Matrix:

[[38 12]
 [10 40]]


Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.76      0.78        50
           1       0.77      0.80      0.78        50

    accuracy                           0.78       100
   macro avg       0.78      0.78      0.78       100
weighted avg       0.78      0.78      0.78       100

Accuracy: 0.78
F1_score: 0.7799119647859143
ROC-AUC score: 0.78
```

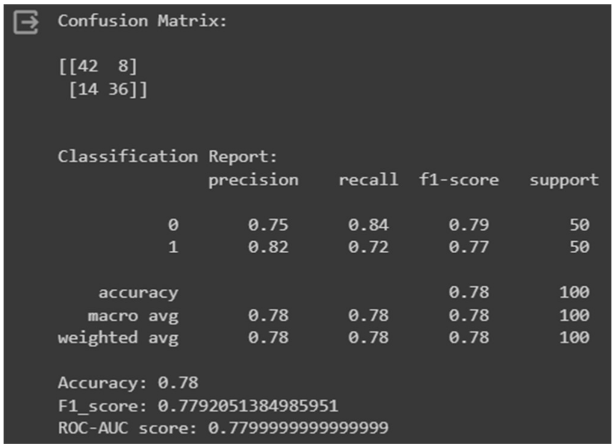


The model correctly predicted 76% of the negative sentiments correctly from the actual negative sentiments and 80% of the positive sentiments correctly from the actual positive ones.

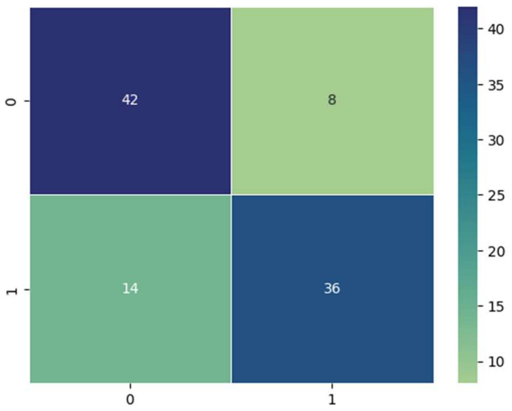


2.3. SUPPORT VECTOR CLASSIFIER

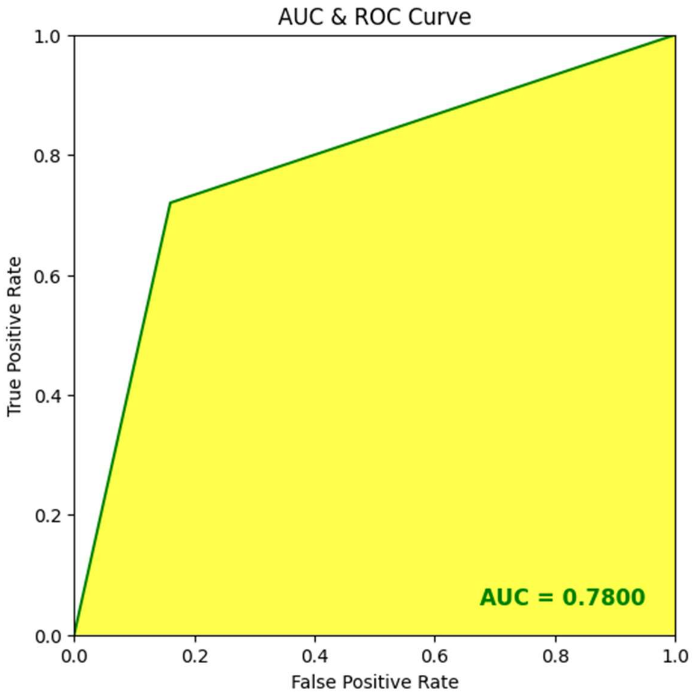
SVC with linear kernel application matched MNB in accuracy, achieving 78%. With a confusion matrix showing 42 true negatives, 36 true positives, 8 false positives, and 14 false negatives, the model performed in a balanced manner. The trade-off between precision and recall as shown below demonstrated strong classification abilities, highlighting the model's dependability in identifying different emotions.



```
Confusion Matrix:

[[42  8]
 [14 36]]


Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.84      0.79        50
           1       0.82      0.72      0.77        50

    accuracy                           0.78       100
   macro avg       0.78      0.78      0.78       100
weighted avg       0.78      0.78      0.78       100

Accuracy: 0.78
F1_score: 0.7792051384985951
ROC-AUC score: 0.7799999999999999
```

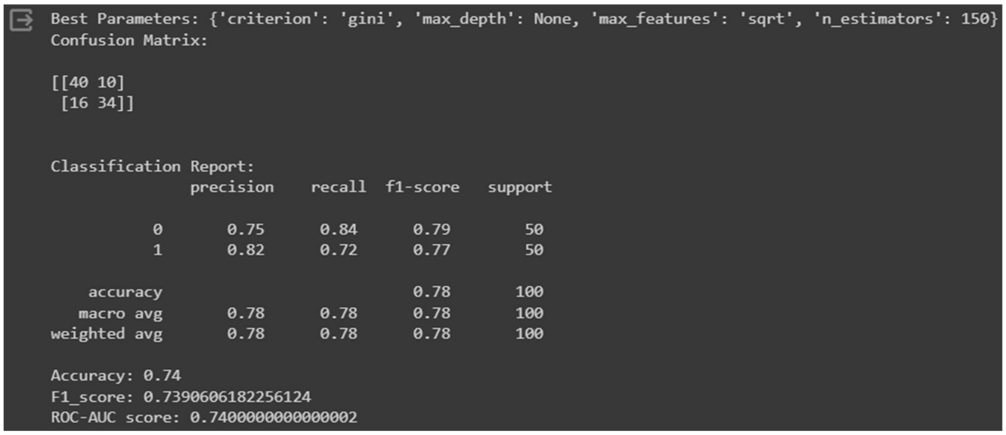


With 84% of the negative sentiments correctly predicted from the actual negative sentiments and 72% of the positive sentiments correctly predicted from the actual positive ones, the model correctly predicted more negative sentiments than positive ones. In contrast, the previous MNB model correctly predicted most of the positive sentiments.



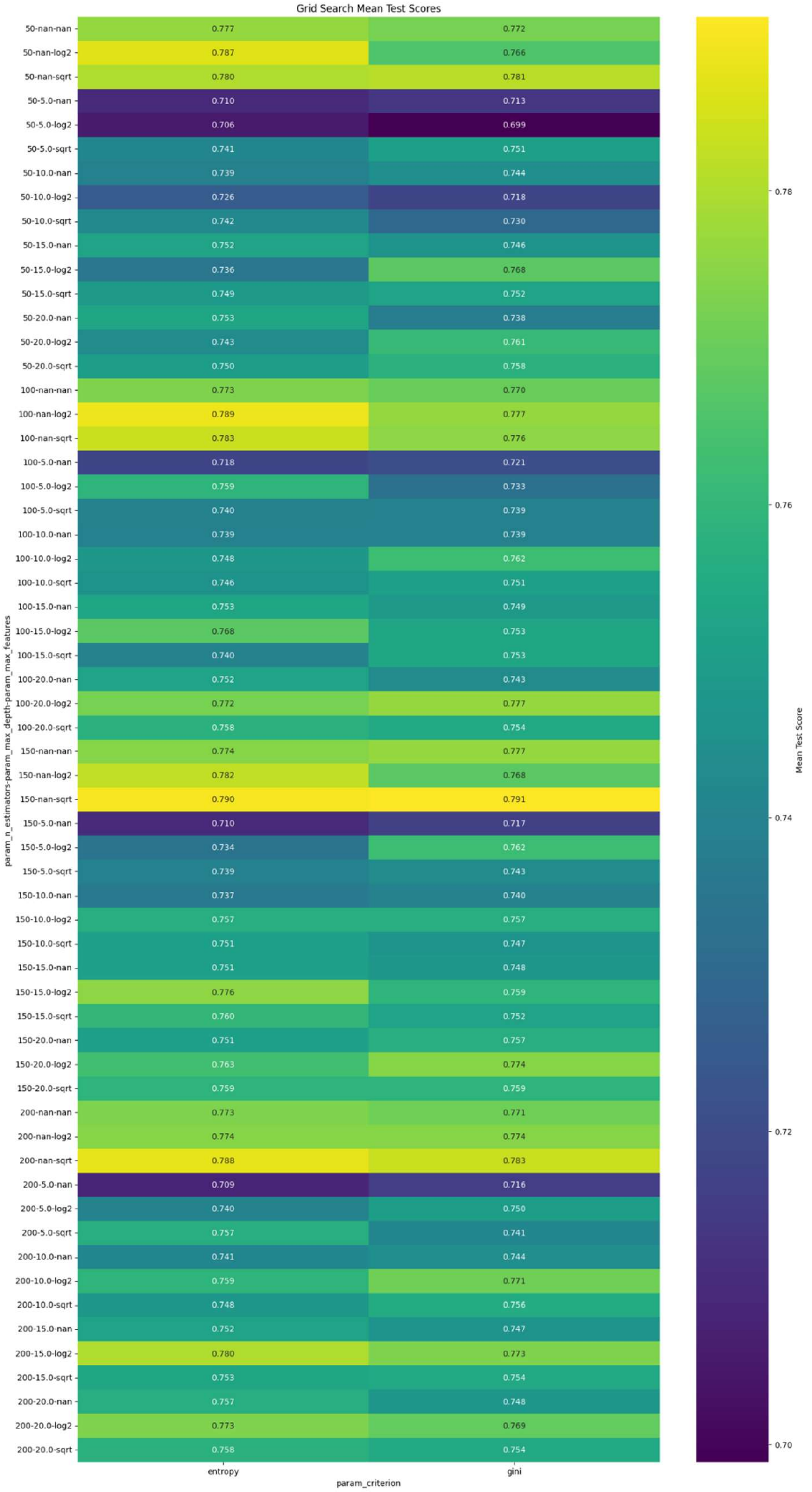
2.4. RANDOM FOREST MODEL CLASSIFIER

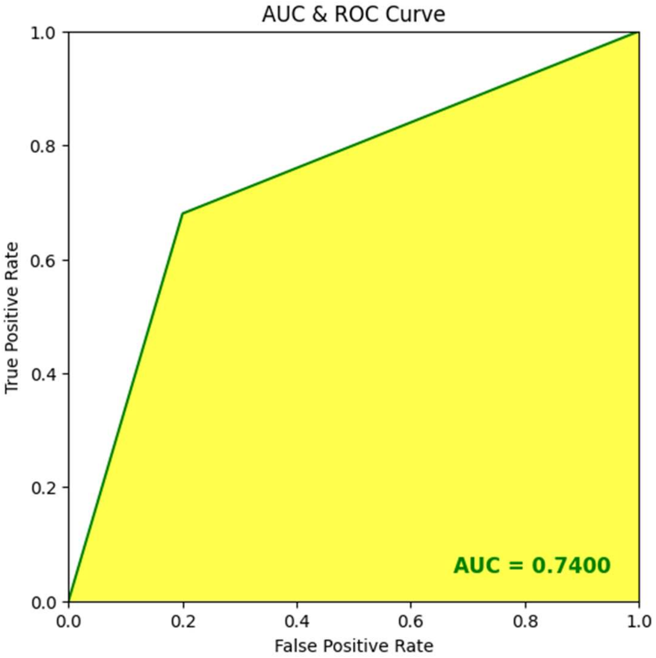
When Random Forest was integrated, accuracy slightly decreased to 74%. The best hyperparameters were found via grid search to be {'criterion': 'gini','max_depth': None,'max_features':'sqrt', 'n_estimators': 150}. There were 40 true negatives, 34 true positives, 10 false positives, and 16 false negatives, according to the confusion matrix. The model continued to exhibit excellent accuracy, but a careful examination of precision and recall provided important new information about the model's classification patterns.



```
Best Parameters: {'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'n_estimators': 150}
Confusion Matrix:

[[40 10]
 [16 34]]


Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.84      0.79        50
           1       0.82      0.72      0.77        50

    accuracy                           0.78       100
   macro avg       0.78      0.78      0.78       100
weighted avg       0.78      0.78      0.78       100

Accuracy: 0.74
F1_score: 0.7390606182256124
ROC-AUC score: 0.7400000000000002
```

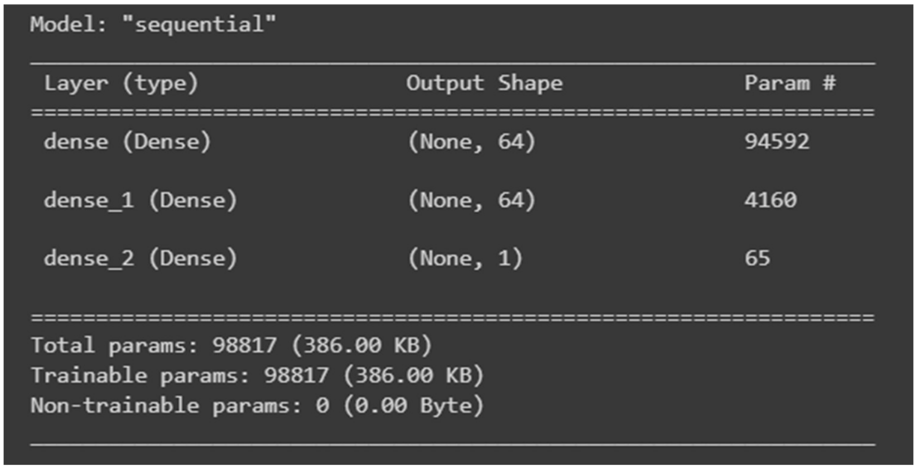
Below is the outcome of the grid search that shows each hyperparameter combination's cross validation accuracy.





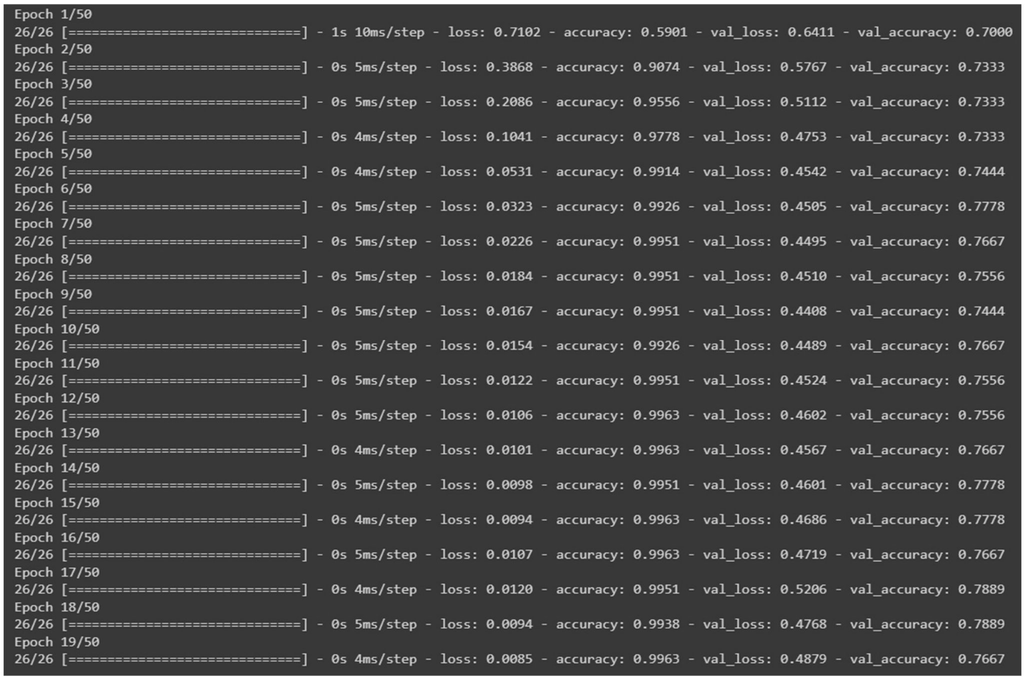
2.5. ANN SEQUENTIAL MODEL

An ANN was used to add an additional layer of analysis. The model outperformed conventional machine learning methods, displaying a competitive accuracy of 80%. Using the 'relu' activation function, the model was trained using a sequential architecture with two hidden layers, each containing 64 neurons. Overfitting during training was avoided by the early stopping mechanism. The summary of the model is shown below.

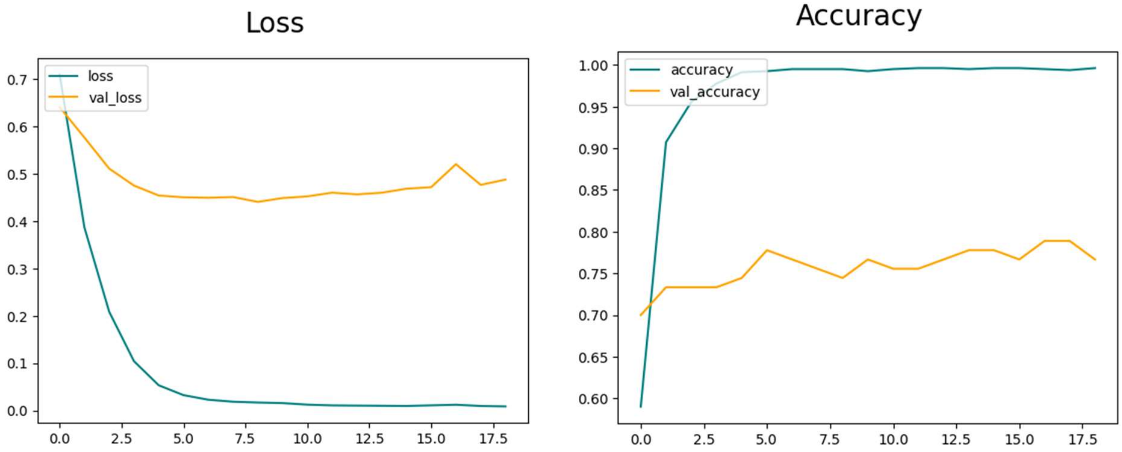


```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                94592

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 98817 (386.00 KB)
Trainable params: 98817 (386.00 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

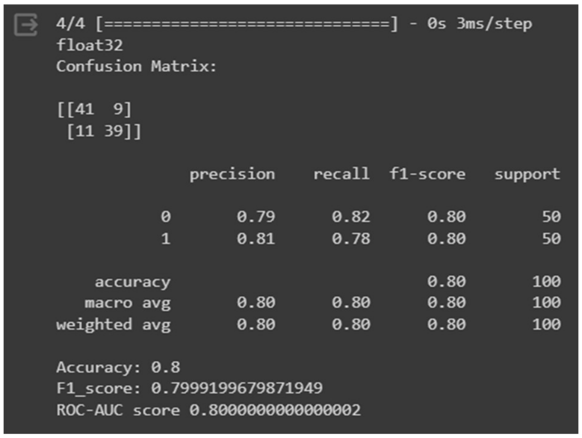
The following plot displays the accuracy during training as well as the loss during validation. As we can see, once the validation loss remains constant for ten consecutive epochs, the model has ceased training after the nineteenth epoch.



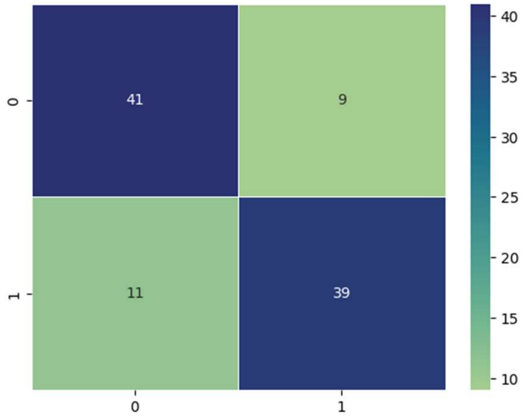
```
Epoch 1/50
26/26 [==============================] - 1s 10ms/step - loss: 0.7102 - accuracy: 0.5901 - val_loss: 0.6411 - val_accuracy: 0.7000
Epoch 2/50
26/26 [==============================] - 0s 5ms/step - loss: 0.3868 - accuracy: 0.9074 - val_loss: 0.5767 - val_accuracy: 0.7333
Epoch 3/50
26/26 [==============================] - 0s 5ms/step - loss: 0.2086 - accuracy: 0.9556 - val_loss: 0.5112 - val_accuracy: 0.7333
Epoch 4/50
26/26 [==============================] - 0s 4ms/step - loss: 0.1041 - accuracy: 0.9778 - val_loss: 0.4753 - val_accuracy: 0.7333
Epoch 5/50
26/26 [==============================] - 0s 4ms/step - loss: 0.0531 - accuracy: 0.9914 - val_loss: 0.4542 - val_accuracy: 0.7444
Epoch 6/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0323 - accuracy: 0.9926 - val_loss: 0.4505 - val_accuracy: 0.7778
Epoch 7/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0226 - accuracy: 0.9951 - val_loss: 0.4495 - val_accuracy: 0.7667
Epoch 8/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0184 - accuracy: 0.9951 - val_loss: 0.4510 - val_accuracy: 0.7556
Epoch 9/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0167 - accuracy: 0.9951 - val_loss: 0.4408 - val_accuracy: 0.7444
Epoch 10/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0154 - accuracy: 0.9926 - val_loss: 0.4489 - val_accuracy: 0.7667
Epoch 11/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0122 - accuracy: 0.9951 - val_loss: 0.4524 - val_accuracy: 0.7556
Epoch 12/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0106 - accuracy: 0.9963 - val_loss: 0.4602 - val_accuracy: 0.7556
Epoch 13/50
26/26 [==============================] - 0s 4ms/step - loss: 0.0101 - accuracy: 0.9963 - val_loss: 0.4567 - val_accuracy: 0.7667
Epoch 14/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0098 - accuracy: 0.9951 - val_loss: 0.4601 - val_accuracy: 0.7778
Epoch 15/50
26/26 [==============================] - 0s 4ms/step - loss: 0.0094 - accuracy: 0.9963 - val_loss: 0.4686 - val_accuracy: 0.7778
Epoch 16/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0107 - accuracy: 0.9963 - val_loss: 0.4719 - val_accuracy: 0.7667
Epoch 17/50
26/26 [==============================] - 0s 4ms/step - loss: 0.0120 - accuracy: 0.9951 - val_loss: 0.5206 - val_accuracy: 0.7889
Epoch 18/50
26/26 [==============================] - 0s 5ms/step - loss: 0.0094 - accuracy: 0.9938 - val_loss: 0.4768 - val_accuracy: 0.7889
Epoch 19/50
26/26 [==============================] - 0s 4ms/step - loss: 0.0085 - accuracy: 0.9963 - val_loss: 0.4879 - val_accuracy: 0.7667
```

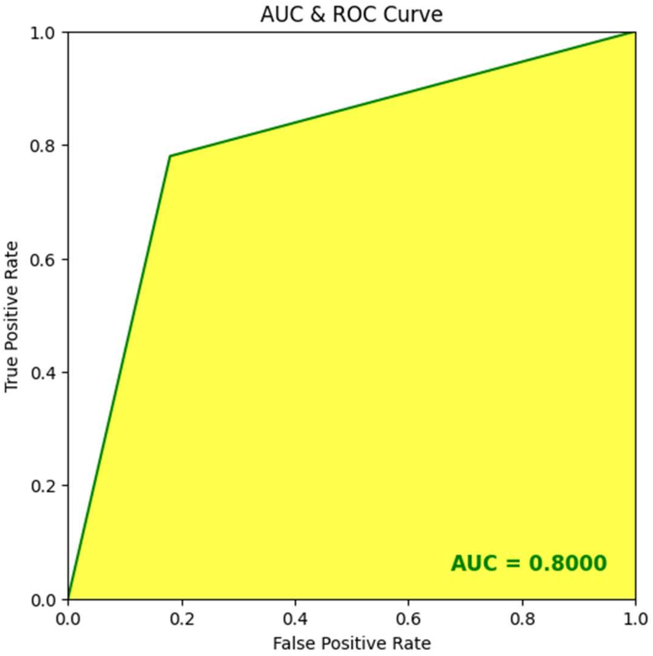


The validation loss of the model stays constant after a few epochs, while the training loss continues to decrease. This suggests that the model's architecture is too simplistic to enable the model to learn additional patterns from the data. A higher performance can be obtained for the model by adding more layers or neurons, which increases the model's complexity. Below are the outcomes of the model's performance using the test set.



```
4/4 [==============================] - 0s 3ms/step
float32
Confusion Matrix:

[[41  9]
 [11 39]]

              precision    recall  f1-score   support

           0       0.79      0.82      0.80        50
           1       0.81      0.78      0.80        50

    accuracy                           0.80       100
   macro avg       0.80      0.80      0.80       100
weighted avg       0.80      0.80      0.80       100

Accuracy: 0.8
F1_score: 0.7999199679871949
ROC-AUC score 0.8000000000000002
```





Out of all the models, the ANN had the best accuracy, coming in at 80%. This suggests that eighty percent of the model's predictions match the test dataset's real sentiments.

By striking a balance between recall and precision, the ANN demonstrated its capacity to precisely forecast positive emotions without compromising its capacity to record all positive occurrences.

2.6. COMPARATIVE ANALYSIS

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC SCORE
Multinomial Naïve Bayes	78%	Balanced	Balanced	Balanced	Balanced
Support Vector Classifier	78%	**High**	Slightly Lower	Balanced	Balanced
Random Forest Classifier	74%	Balanced	Balanced	Balanced	Balanced
ANN	**80%**	Balanced	Balanced	**High**	**High**

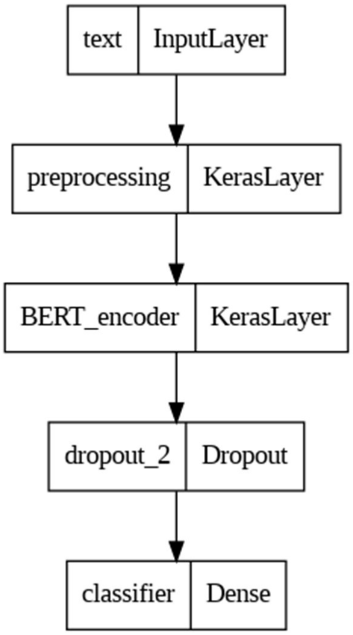
A useful way to understand the relative strengths of MNB, SVC, RFC, and ANN are to compare their performance metrics. Although the accuracy obtained by MNB and SVC was comparable, there are minor differences when precision, recall, and the F1 score are carefully analysed. While SVC showed slightly higher precision at the expense of recall, MNB demonstrated a balance between precision and recall. RFC preserved a balanced precision-recall trade-off despite being slightly less accurate. With its neural architecture, the ANN showed the best accuracy of all the models, indicating that it can be used to capture complex sentiment patterns.

3. CLASSIFICATION USING SMALL BERT MODEL

3.1. OVERVIEW

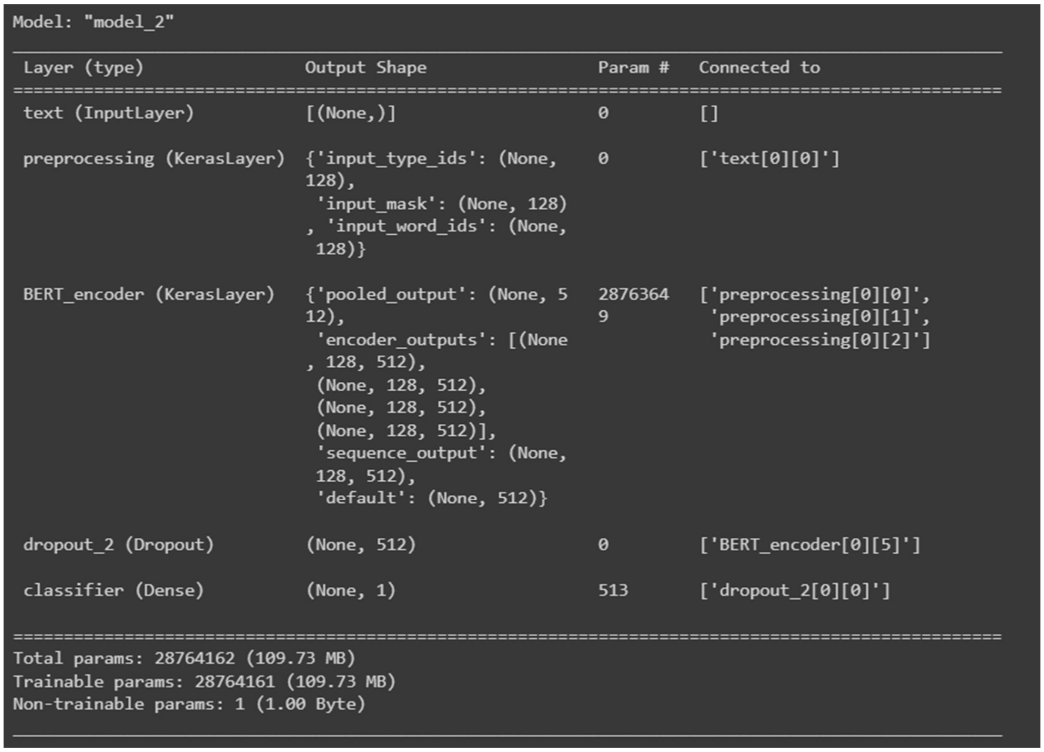
Importing the required libraries and defining the preprocessing functions is where the code starts. Subsequently, it applies text classification using the Small BERT model (bert_en_uncased_L-4_H-512_A-8/1) from TensorFlow Hub. Adam serves as the optimizer and binary cross entropy as the loss function when compiling the neural network.

MODEL ARCHITECTURE



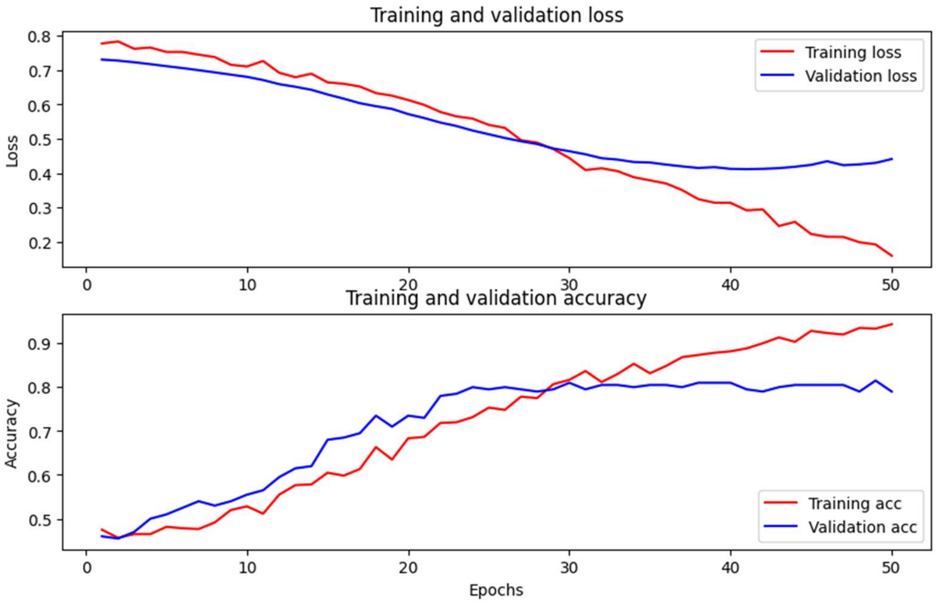
3.2. FINE-TUNING OF MODEL

A dense layer of one neuron is added to the model to fine-tune it. For the binary classification task, this layer activates in a "sigmoid" manner, and the pre-trained weights are set to update as the model is trained. The model’s summary is shown below.



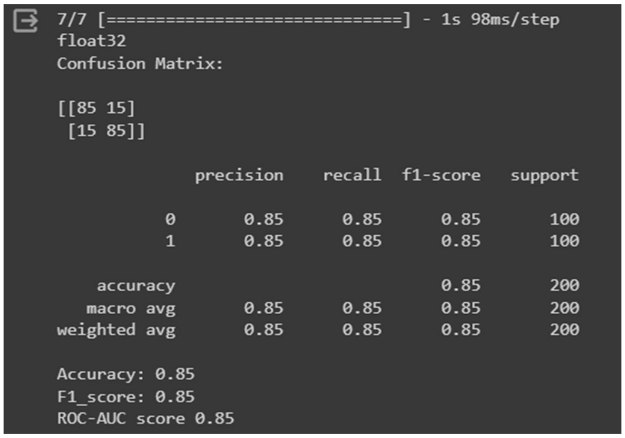
```
Model: "model_2"
__________________________________________________________________________________________________
 Layer (type)                Output Shape                 Param #   Connected to
==================================================================================================
 text (InputLayer)           [(None,)]                    0         []

 preprocessing (KerasLayer)  {'input_type_ids': (None,    0         ['text[0][0]']
                             128),
                              'input_mask': (None, 128)
                             , 'input_word_ids': (None,
                              128)}

 BERT_encoder (KerasLayer)   {'pooled_output': (None, 5   2876364   ['preprocessing[0][0]',
                             12),                         9          'preprocessing[0][1]',
                              'encoder_outputs': [(None              'preprocessing[0][2]']
                             , 128, 512),
                              (None, 128, 512),
                              (None, 128, 512),
                              (None, 128, 512)],
                              'sequence_output': (None,
                              128, 512),
                              'default': (None, 512)}

 dropout_2 (Dropout)         (None, 512)                  0         ['BERT_encoder[0][5]']

 classifier (Dense)          (None, 1)                    513       ['dropout_2[0][0]']

==================================================================================================
Total params: 28764162 (109.73 MB)
Trainable params: 28764161 (109.73 MB)
Non-trainable params: 1 (1.00 Byte)
__________________________________________________________________________________________________
```

The model performs well after 50 epochs of training on the given dataset. The validation set's binary accuracy hits 81.5%, demonstrating the model's capacity to generalise and function well with unknown data. Effective learning is indicated by a steady decrease in loss during training. The loss and accuracy of the training and validation sets, as we can see, begin to diverge after the 40th epoch, indicating that the model began overfitting the training set.

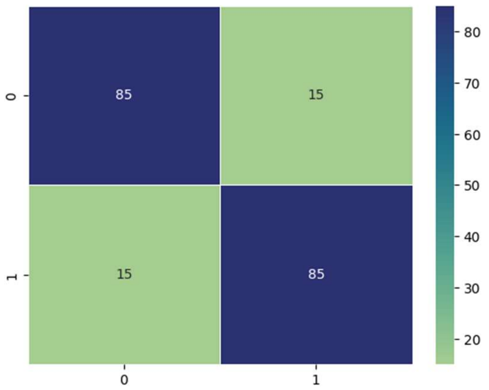


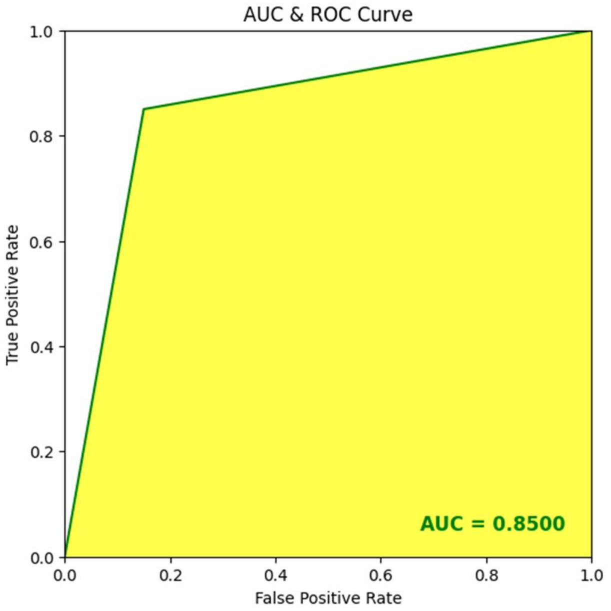
3.3. PERFORMANCE ANALYSIS

The assessment metrics offer a thorough comprehension of the model's functionality. For both classes (0 and 1), the confusion matrix displays balanced precision and recall. With an accuracy, F1-score, and ROC-AUC score all close to 85%, the model appears to be functioning well.



```
7/7 [==============================] - 1s 98ms/step
float32
Confusion Matrix:

[[85 15]
 [15 85]]

              precision    recall  f1-score   support

           0       0.85      0.85      0.85       100
           1       0.85      0.85      0.85       100

    accuracy                           0.85       200
   macro avg       0.85      0.85      0.85       200
weighted avg       0.85      0.85      0.85       200

Accuracy: 0.85
F1_score: 0.85
ROC-AUC score 0.85
```





The trade-off between true positive rate and false positive rate is graphically represented by the ROC-AUC curve. The model's discriminative power is reinforced by the AUC value of 0.85, and the curve exhibits a good balance.

3.4. EVALUTION OF PERFORMANCE COMPARED TO EARLIER MODELS

Interesting insights are revealed when this model is compared to the four previously trained models in previous task. All models show competitive performance; however, in terms of accuracy and AUC, the Small BERT model appears to be slightly superior. The F1-score, precision, and recall metrics are in good agreement with the other models.

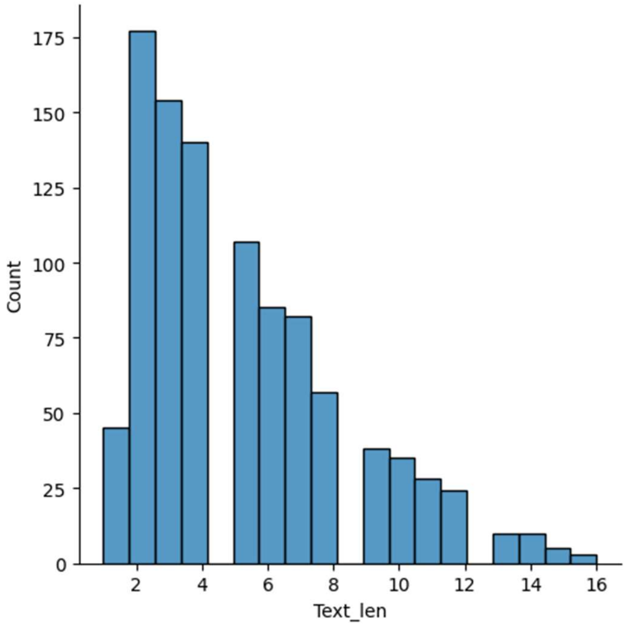
To sum up, the Small BERT model that was trained for text classification shows strong learning capabilities and produces excellent results on the given dataset. Its performance is comparable to that of the other models trained in task 2, if not slightly superior. The analysis emphasises how pre-trained models can be used effectively for natural language processing tasks.

4. TOPIC DETECTION

4.1. TOPIC DETECTION USING BERTOPIC

4.1.1. MODEL TRAINING

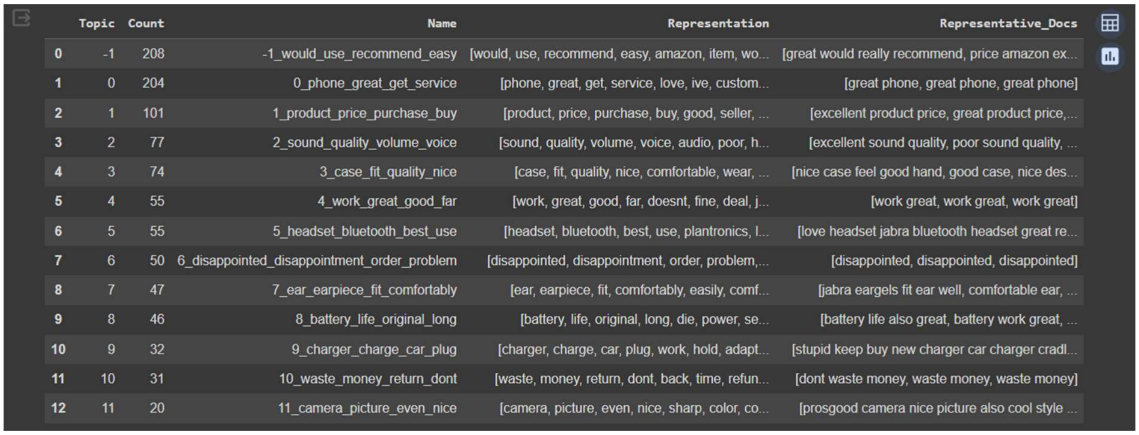
The first step is to comprehend the sentence length distribution, which shows that the longest review is sixteen words long. Words with two or four letters have the highest count. This indicates that the language employed in our dataset is not overly complex.



After employing the "paraphrase-MiniLM-L3-v2" embedding model to train the BERTopic model, 13 topics were identified. A summary of the various themes found in the dataset is provided for each topic through representative words and documents.

4.1.2. RESULTS

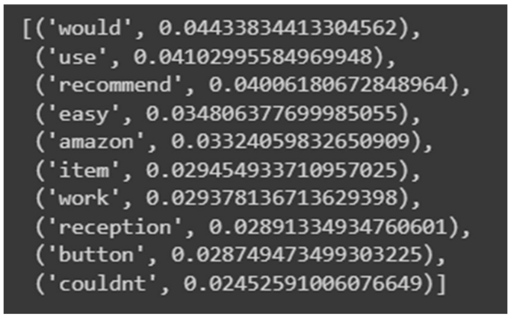
The 'get_topic_info()' method of the model provides a more comprehensive overview of all the topics, complete with counts, names, representative words, and sample sentences.



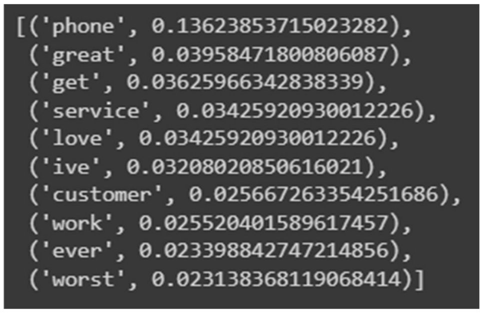
	Topic	Count	Name	Representation	Representative_Docs
0	-1	208	-1_would_use_recommend_easy	[would, use, recommend, easy, amazon, item, wo...	[great would really recommend, price amazon ex...
1	0	204	0_phone_great_get_service	[phone, great, get, service, love, ive, custom...	[great phone, great phone, great phone]
2	1	101	1_product_price_purchase_buy	[product, price, purchase, buy, good, seller, ...	[excellent product price, great product price,...
3	2	77	2_sound_quality_volume_voice	[sound, quality, volume, voice, audio, poor, h...	[excellent sound quality, poor sound quality, ...
4	3	74	3_case_fit_quality_nice	[case, fit, quality, nice, comfortable, wear, ...	[nice case feel good hand, good case, nice des...
5	4	55	4_work_great_good_far	[work, great, good, far, doesnt, fine, deal, j...	[work great, work great, work great]
6	5	55	5_headset_bluetooth_best_use	[headset, bluetooth, best, use, plantronics, l...	[love headset jabra bluetooth headset great re...
7	6	50	6_disappointed_disappointment_order_problem	[disappointed, disappointment, order, problem,...	[disappointed, disappointed, disappointed]
8	7	47	7_ear_earpiece_fit_comfortably	[ear, earpiece, fit, comfortably, easily, comf...	[jabra eargels fit ear well, comfortable ear, ...
9	8	46	8_battery_life_original_long	[battery, life, original, long, die, power, se...	[battery life also great, battery work great, ...
10	9	32	9_charger_charge_car_plug	[charger, charge, car, plug, work, hold, adapt...	[stupid keep buy new charger car charger cradl...
11	10	31	10_waste_money_return_dont	[waste, money, return, dont, back, time, refun...	[dont waste money, waste money, waste money]
12	11	20	11_camera_picture_even_nice	[camera, picture, even, nice, sharp, color, co...	[prosgood camera nice picture also cool style ...

Based on representative words, a unique label is assigned to each topic. The number of sentences connected to each topic is shown in the Count column. A wide range of topics are covered, such as suggested products, phone quality, cost, fit, sound, and customer support. Exemplary documents provide instances to help comprehend every subject.

The c-TF-IDF (cluster Term Frequency-Inverse Document Frequency) of each term within a cluster document is also displayed in the below figures. Higher scores indicate the key terms that define a particular topic.



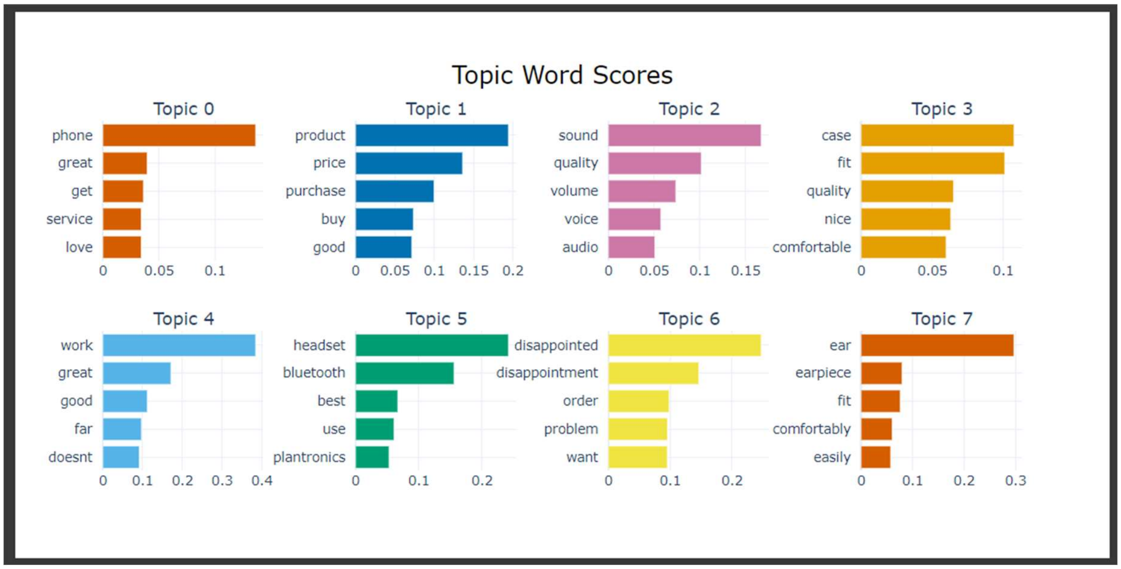
```
[('would', 0.04433834413304562),
 ('use', 0.04102995584969948),
 ('recommend', 0.04006180672848964),
 ('easy', 0.03480637699985055),
 ('amazon', 0.03324059832650909),
 ('item', 0.029454933710957025),
 ('work', 0.029378136713629398),
 ('reception', 0.02891334934760601),
 ('button', 0.028749473499303225),
 ('couldnt', 0.02452591006076649)]
```

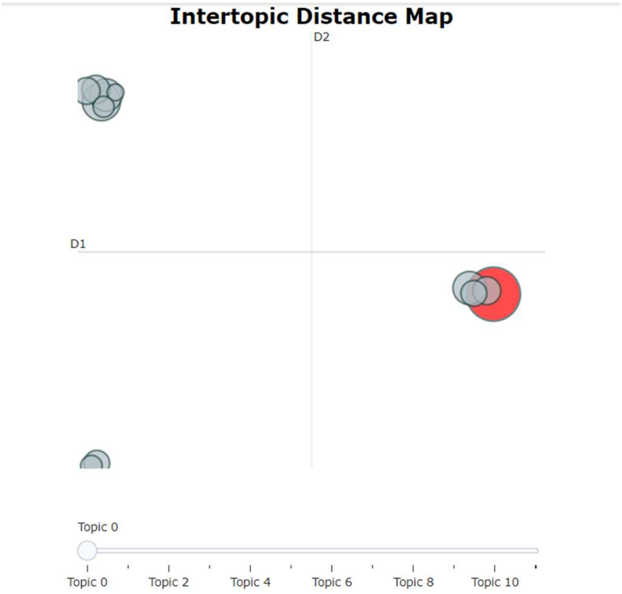


```
[('phone', 0.13623853715023282),
 ('great', 0.03958471800806087),
 ('get', 0.03625966342838339),
 ('service', 0.03425920930012226),
 ('love', 0.03425920930012226),
 ('ive', 0.03208020850616021),
 ('customer', 0.025667263354251686),
 ('work', 0.025520401589617457),
 ('ever', 0.023398842747214856),
 ('worst', 0.023138368119068414)]
```

The initial c-TF-IDF depiction suggests that the "Recommendations and Ease of Use" subject matter highlights expressions like "would," "use," and "recommend," with notable emphasis placed on "easy" and "amazon." This shows that using Amazon products is simple and that they come highly recommended.

The terms "phone," "great," and "service," which convey positive opinions about phone features and services, are used to highlight the "Phone Reviews" topic in the second representation. The terms "love" and "customer" are prominent, indicating positive user experiences, whereas "worst" indicates unfavourable reviews.





4.1.3. ANALYSIS OF TOPICS

Recommendations and Ease of Use: Contributors in this area share positive opinions, highlighting the usability of products—especially those from Amazon—and offering positive feedback. 'Recommend' and 'easy' are important terms in these conversations.

Phone Reviews: Users who are satisfied with the features and services of their phones leave positive reviews on this topic. "Great" and "love" are important terms that denote a satisfying user experience.

Product Price and Purchase: This category includes discussions about the cost of products, where to buy them, and their general quality. Users evaluate value for money, noting factors such as "good" and "buy" when making decisions about purchases.

Sound Quality and Volume: Users highlight voice clarity, volume levels, and sound quality when assessing sound-related factors. Phrases such as 'good' and 'bad' offer insights into their evaluations.

Case Features and Quality: In this discussion, users discuss the fit, comfort, and general quality of cases as well as their features. Important words that emphasise important factors are "nice" and "comfortable."

Work Performance: Users talk about how well products work in professional environments, highlighting satisfying experiences. The adjectives "excellent" and "good" denote contentment with the functionalities linked to the work.

Headset and Bluetooth Devices: Users discuss headsets and Bluetooth devices in this topic, expressing preference for certain brands such as Plantronics. Positive sentiments are indicated by key terms such as "love" and "great."

Disappointment and Order Problems: Users in this group talk about problems with their orders and express disappointment. Words like "problem" and "disappointed" draw attention to unfavourable experiences and possible areas for development.

Fit and Comfort of Earpieces: Discussions on the topic of earpiece fit and comfort feature user feedback on both general comfort and ease of fitting. The word "comfortably" denotes an emphasis on ergonomic factors.

Battery Life: Discussions in this category centre on the longevity and uniqueness of batteries, with users evaluating both. Important phrases like "long" and "die" provide information about their assessments.

Charger Issues: Discussions about problems and encounters with chargers, including car chargers and plug features, are found here. Words like "plug" and "work" shed light on their evaluations.

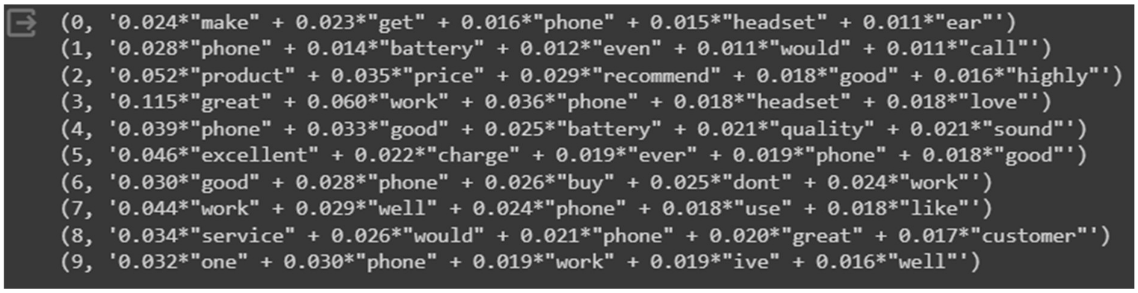
Waste of Money: Discussions where users voice their displeasure and call out specific products as a waste of money are captured under this topic. Words like "waste" and "return" imply unfavourable feelings and possible issues.

Camera and Picture Quality: Users comment on clarity and aesthetics while focusing on camera and picture quality. Qualities such as 'nice' and 'sharp' offer insights into their evaluations.

4.2. TOPIC DETECTION USING LATENT DIRICHLET ALLOCATION (LDA)

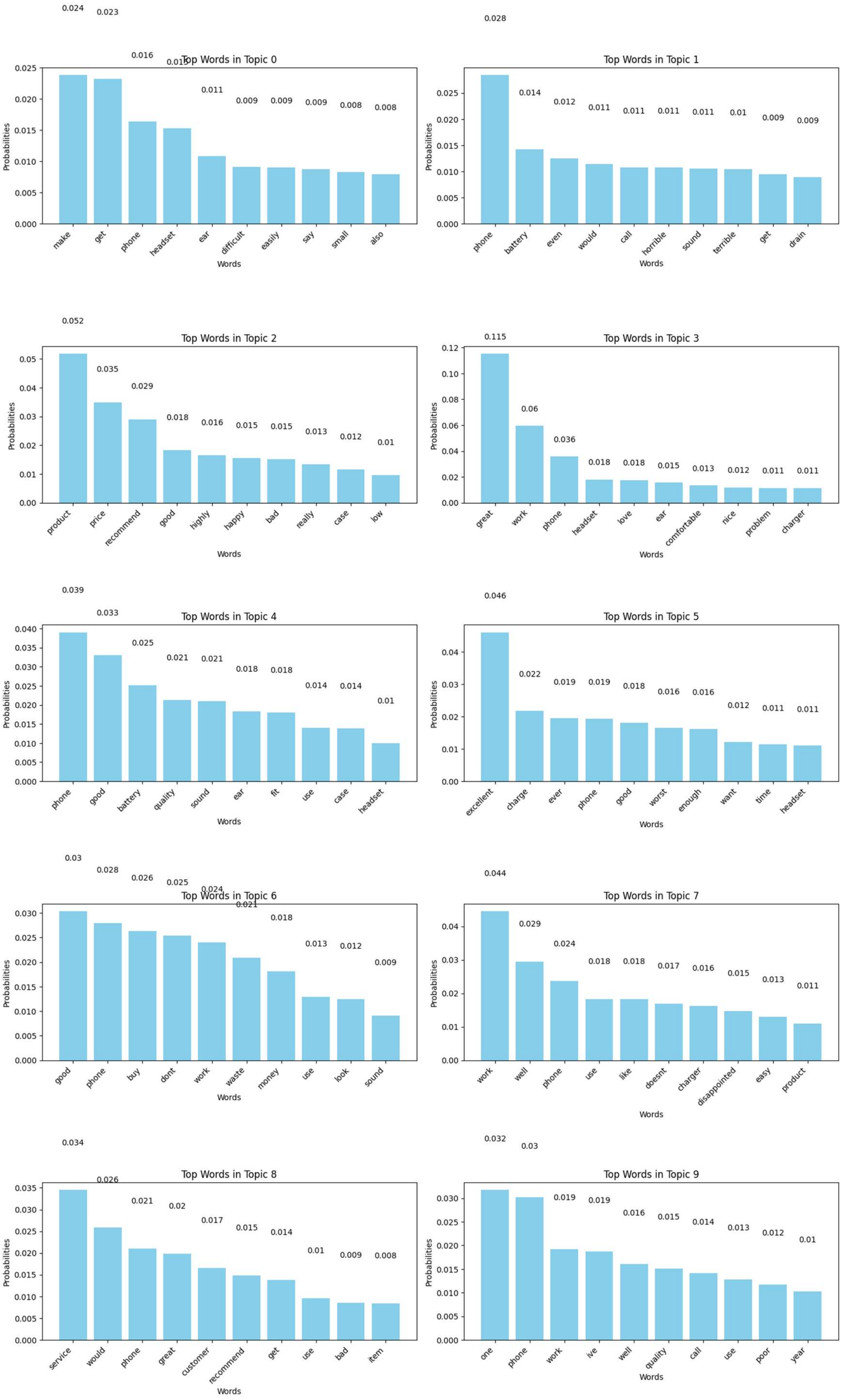
4.2.1. MODEL RESULTS

Different discussions are found in the dataset according to the LDA topic modelling analysis.10 topics were identified such as general device use, phone activities, product recommendations, satisfying device experiences, evaluations of device quality, charging experiences, purchase decisions, device functionality, feedback from customer service, and individual device experiences. Users share their opinions on a variety of device-related topics, offering insightful commentary on their experiences as well as suggestions and worries.



```
(0, '0.024*"make" + 0.023*"get" + 0.016*"phone" + 0.015*"headset" + 0.011*"ear"')
(1, '0.028*"phone" + 0.014*"battery" + 0.012*"even" + 0.011*"would" + 0.011*"call"')
(2, '0.052*"product" + 0.035*"price" + 0.029*"recommend" + 0.018*"good" + 0.016*"highly"')
(3, '0.115*"great" + 0.060*"work" + 0.036*"phone" + 0.018*"headset" + 0.018*"love"')
(4, '0.039*"phone" + 0.033*"good" + 0.025*"battery" + 0.021*"quality" + 0.021*"sound"')
(5, '0.046*"excellent" + 0.022*"charge" + 0.019*"ever" + 0.019*"phone" + 0.018*"good"')
(6, '0.030*"good" + 0.028*"phone" + 0.026*"buy" + 0.025*"dont" + 0.024*"work"')
(7, '0.044*"work" + 0.029*"well" + 0.024*"phone" + 0.018*"use" + 0.018*"like"')
(8, '0.034*"service" + 0.026*"would" + 0.021*"phone" + 0.020*"great" + 0.017*"customer"')
(9, '0.032*"one" + 0.030*"phone" + 0.019*"work" + 0.019*"ive" + 0.016*"well"')
```

The below graph shows the probability distributions of each word in the topic.



Top Words in Topic 0
Probabilities
Words
0.000
0.005
0.010
0.015
0.020
0.025
make
get
phone
headset
ear
difficult
easily
say
small
also
0.024
0.023
0.016
0.015
0.011
0.009
0.009
0.009
0.008
0.008
Top Words in Topic 1
Probabilities
Words
0.000
0.005
0.010
0.015
0.020
0.025
phone
battery
even
would
call
horrible
sound
terrible
get
drain
0.028
0.014
0.012
0.011
0.011
0.011
0.011
0.01
0.009
0.009
Top Words in Topic 2
Probabilities
Words
0.00
0.01
0.02
0.03
0.04
0.05
product
price
recommend
good
highly
happy
bad
really
case
low
0.052
0.035
0.029
0.018
0.016
0.015
0.015
0.013
0.012
0.01
Top Words in Topic 3
Probabilities
Words
0.00
0.02
0.04
0.06
0.08
0.10
0.12
great
work
phone
headset
love
ear
comfortable
nice
problem
charger
0.115
0.06
0.036
0.018
0.018
0.015
0.013
0.012
0.011
0.011
Top Words in Topic 4
Probabilities
Words
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
phone
good
battery
quality
sound
ear
fit
use
case
headset
0.039
0.033
0.025
0.021
0.021
0.018
0.018
0.014
0.014
0.01
Top Words in Topic 5
Probabilities
Words
0.00
0.01
0.02
0.03
0.04
excellent
charge
ever
phone
good
worst
enough
want
time
headset
0.046
0.022
0.019
0.019
0.018
0.016
0.016
0.012
0.011
0.011
Top Words in Topic 6
Probabilities
Words
0.000
0.005
0.010
0.015
0.020
0.025
0.030
good
phone
buy
dont
work
waste
money
use
look
sound
0.03
0.028
0.026
0.025
0.024
0.021
0.018
0.013
0.012
0.009
Top Words in Topic 7
Probabilities
Words
0.00
0.01
0.02
0.03
0.04
work
well
phone
use
like
doesnt
charger
disappointed
easy
product
0.044
0.029
0.024
0.018
0.018
0.017
0.016
0.015
0.013
0.011
Top Words in Topic 8
Probabilities
Words
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
service
would
phone
great
customer
recommend
get
use
bad
item
0.034
0.026
0.021
0.02
0.017
0.015
0.014
0.01
0.009
0.008
Top Words in Topic 9
Probabilities
Words
0.000
0.005
0.010
0.015
0.020
0.025
0.030
one
phone
work
ive
well
quality
call
use
poor
year
0.032
0.03
0.019
0.019
0.016
0.015
0.014
0.013
0.012
0.01

4.2.2. ANALYSIS OF TOPICS

TOPICS	KEYWORDS	INTERPRETATION
1.	make, get, phone, headset, ear	General device use, including phone calls and headset use. The terms "phone" and "headset" are prominent, indicating different people's opinions about how the device works.
2.	phone, battery, even, would, call	Conversations about different phone activities, such as battery life and call-related incidents. A focus on phone performance and features is indicated by the emphasis on the words "phone" and "battery."
3.	product, price, recommend, good, highly	Topic centred on product recommendations, with users stressing the value of the suggested products' prices and quality. The frequency of "product" and "price" indicates value and recommendation considerations.
4.	great, work, phone, headset, love	Positive device experiences that highlight their greatness, practicality in the workplace, and fondness for devices are expressed. Words like "love" and "great" denote contentment.
5.	phone, good, battery, quality, sound	Pay attention to the device's quality, considering features like sound and battery life. When evaluating a device's features, users often take "quality" and "sound" into account.
6.	excellent, charge, ever, phone, good	Appreciations for the excellent devices and the conversations about the charging process. Good ratings are indicated by the prominence of "excellent," and conversations about charging functionality are indicated by the word "charge."
7.	good, phone, buy, dont, work	Insights on what to buy, considering things like functionality and device quality. Words like "buy" and "dont" imply factors that affect purchasing decisions.
8.	work, well, phone, use, like	Talks focus on the likeability and functionality of devices, highlighting user preferences and how well they function. Words like "like" and "work" denote contentment.
9.	service, would, phone, great, customer	Discussing customer service experiences while taking aspects like service quality into account. The terms 'service' and 'customer' are prominent, indicating that discussions should centre around services.
10.	one, phone, work, ive, well	Conversations concerning the different aspects that affect how devices are used, such as personal preferences and features. Phrases like "work" and "ive" denote judgements from the reviewers.

REFERENCES

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems* (3rd ed.). O'Reilly Media, Inc.

Bastiaan Sjardin, Luca Massaron, & Boschetti, A. (2016). *Large Scale Machine Learning with Python*. Packt Publishing Ltd.