Intelligent Data and Text Analysis Coursework 2

Table of contents

Part 1: Preprocessing

Textual data preprocessing is a critical step in preparing raw text for machine learning and data analysis. This phase involves removing punctuation, numbers, and stop words, as well as lowercasing and lemmatizing words. These techniques collectively eliminate noise, ensuring a cleaner dataset that enhances the effectiveness of subsequent analyses and models.

Punctuation removal:

Punctuation removal is the process of eliminating punctuation marks (such as commas, periods, exclamation marks, etc.) from a piece of text. This preprocessing step simplifies and standardises text, optimising it for subsequent analyses and modelling. Figure 1 illustrates the contrast between the text that includes punctuation and the text without punctuation.

Before	After
So there is no way for me to plug it in here in the US unless I go by a converter.	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.	Good case Excellent value
Great for the jawbone.	Great for the jawbone
Tied to charger for conversations lasting more than 45 minutes.MAJOR PROBLEMS!!	Tied to charger for conversations lasting more than 45 minutesMAJOR PROBLEMS
The mic is great.	The mic is great

Figure 1: Punctuation Removal

In Figure 1, the impact of punctuation removal is clearly illustrated. The dots at the end of each line have been eliminated, and the exclamation mark in line 4, as well as the comma in line 2, have been successfully removed with the application of the punctuation removal process.

Number removal:

Number removal involves eliminating numeric characters (0-9) from text, streamlining and standardising the data for analysis. This preprocessing step reduces noise and ensures a more focused and consistent

representation, optimising the text for subsequent natural language processing tasks.The visual comparison,in Figure 2 for punctuation removal, can effectively showcase the impact of number removal on the original text.

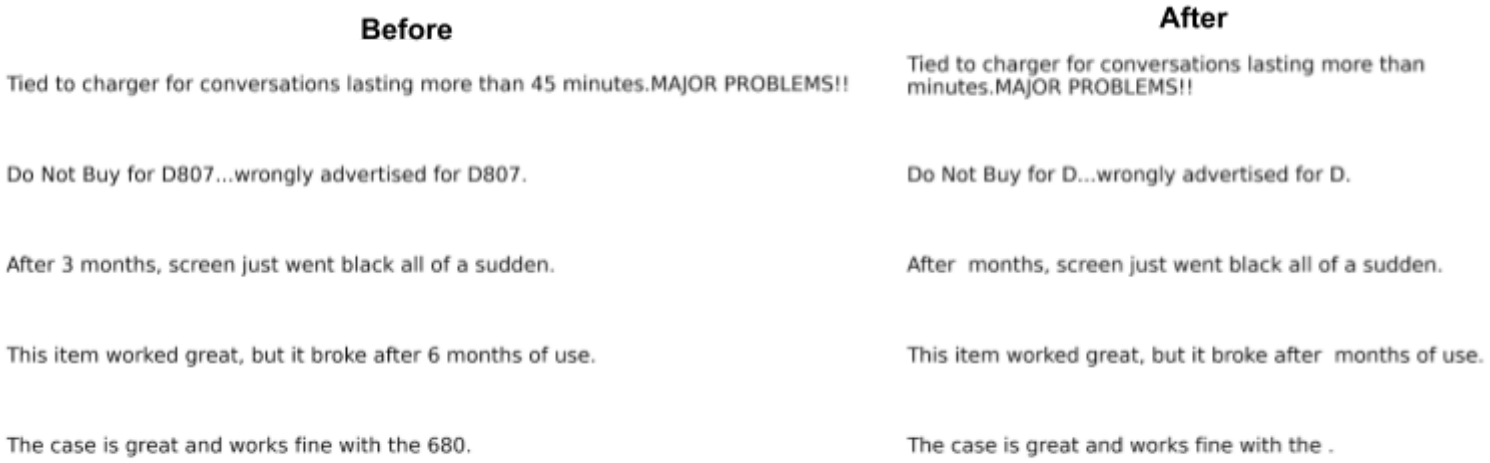

Figure 2: Number removal

In Figure 2, the transformative effect of number removal is evident, where all numeric characters have been successfully eliminated from lines 1 to 5. Notably, specific numbers, such as 45 in line 1, 807 in line 2, 3 in line 3, 6 in line 4, and 680 in line 5.

Lowercasing:

Lowercasing involves converting all letters in text to lowercase, ensuring consistency and simplifying word representation for analysis and modelling. This preprocessing step, illustrated visually in Figure 3, neutralises case variations, contributing to the effectiveness of natural language processing tasks by fostering a standardised representation of the input data.

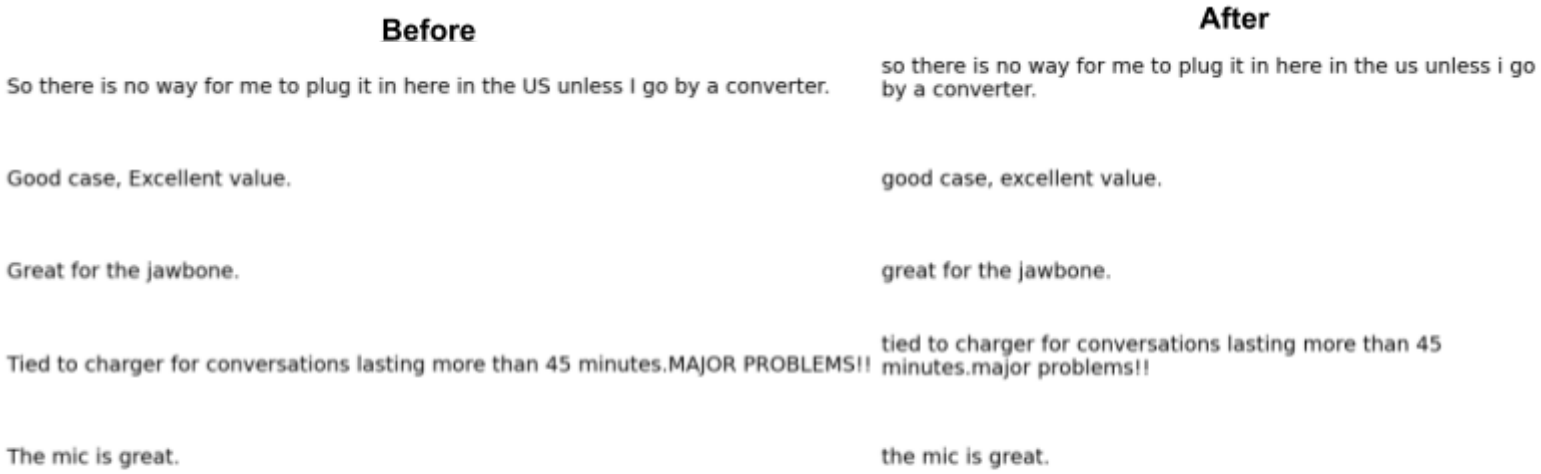

Figure 3: Lowercasing

In Figure 3, the impact of lowercasing is evident, showcasing the transformation of selected words to lowercase. Noteworthy changes include the conversion of "So" to "so" in the first line, "US" to "us,"

"Good" to "good," and "Excellent" to "excellent" in the second line. Additionally, lowercasing has modified "Great" to "great" in the third line, "Tied" to "tied" and "MAJOR PROBLEMS" to "major problems" in the fourth line, and finally, "The" to "the" in the fifth line.

Stopwords Removal:

Stopwords removal is a crucial preprocessing step in text analysis, involving the exclusion of common, non-informative words. This process enhances the text's relevance by focusing on content-rich terms, minimising noise for improved analyses and machine learning tasks. Illustrated in Figure 4, the visual contrast between the original text and the refined version after stopwords removal highlights the impact of this technique on text clarity and relevance.

Figure 4: Stopwords Removal

In Figure 4, the impact of stopwords removal is clear, highlighting a more concise and polished text. Notably, the first line is transformed to "way plug US unless go converter" as common stopwords are eliminated. Line 2, containing no stopwords, remains unchanged. The third line becomes "Great jawbone" by removing the stopwords "for" and "the." Line 4 undergoes changes, turning into "Tied charger conversations lasting 45 minutes. MAJOR PROBLEMS!!" through the removal of stopwords like "to," "for," "more," and "than." Finally, line 5 turns to "mic great" with the exclusion of the stopwords "The" and "is .

Lemmatization:

Lemmatization is a key preprocessing technique in natural language processing, aiming to transform words into their base or root forms. By reducing words to their core forms, lemmatization aids in standardising and simplifying textThis process is particularly valuable in capturing the essential meaning of words and reducing feature dimensionality. Illustrated in Figure 5, the visual representation contrasts the original text with the lemmatized version.

Before	After
So there is no way for me to plug it in here in the US unless I go by a converter.	so there be no way for me to plug it in here in the u unless i go by a converter.
Good case, Excellent value.	good case, excellent value.
Great for the jawbone.	great for the jawbone.
Tied to charger for conversations lasting more than 45 minutes.MAJOR PROBLEMS!!	tie to charger for conversation last more than 45 minutes.major problems!!
The mic is great.	the mic be great.

Figure 5: Lemmatization

In Figure 5, the impact of lemmatization is demonstrated, showcasing the transformation of words to their base or root forms for linguistic simplicity. In the first line, the word "is" was lemmatized to "be." Lines 2 and 3 remained unchanged, as they originally had no lemmatization variations. In the fourth line, lemmatization modified "tied" to "tie," "conversations" to "conversation," and "lasting" to "last," refining the text to its essential linguistic forms. Lastly, in the fifth line, the word "is" underwent lemmatization, transforming into "be".

Part 2: Bag of Words Classification

The reviews were initially shuffled to eliminate any potential bias introduced by the ordering of the data, ensuring an unbiased assessment of the models' performance. Subsequently, the dataset was partitioned into training and testing sets, with 80% of the data allocated to training and 20% to testing. This partitioning ratio is employed to strike a balance between having a sufficiently large training set for model learning and a representative test set for unbiased evaluation. To represent the textual data numerically, both the training and test reviews were transformed using the count vectorization technique. Count vectorization is chosen for its ability to capture the frequency of word occurrences in the text, providing numerical representation for subsequent tasks. Figure 6 shows the comparison of the created models.

Algorithms	Precision	Recall	F1 Score	Accuracy Score	AUC Score
Random Forest	0.8182	0.75	0.7826	0.8	0.7981
Naive Bayes	0.757	0.8438	0.798	0.795	0.7969
Decision Tree	0.8095	0.7083	0.7556	0.78	0.7772
SVM	0.7889	0.7396	0.7634	0.78	0.7784

Figure 6: Models Metric comparison

Decision Tree

The decision tree model exhibited commendable performance in sentiment classification using the bag-of-words representation. With an overall accuracy of 78%, the model demonstrated balanced precision and recall values for both positive and negative sentiments. Specifically, positive sentiment identification achieved a precision of 76% and a recall of 85%, resulting in an F1-score of 80%. Similarly, negative sentiment identification showed a precision of 81%, a recall of 71%, and an F1-score of 76%. The area under the ROC curve (ROC AUC) further confirmed the model's effectiveness, yielding a score of 0.777. The ROC curve is visualised in Figure 8, while the confusion matrix, providing a detailed breakdown of model predictions, can be observed in Figure 7. These results collectively suggest that the decision tree model, trained on bag-of-words representation, is a robust classifier for sentiment analysis.

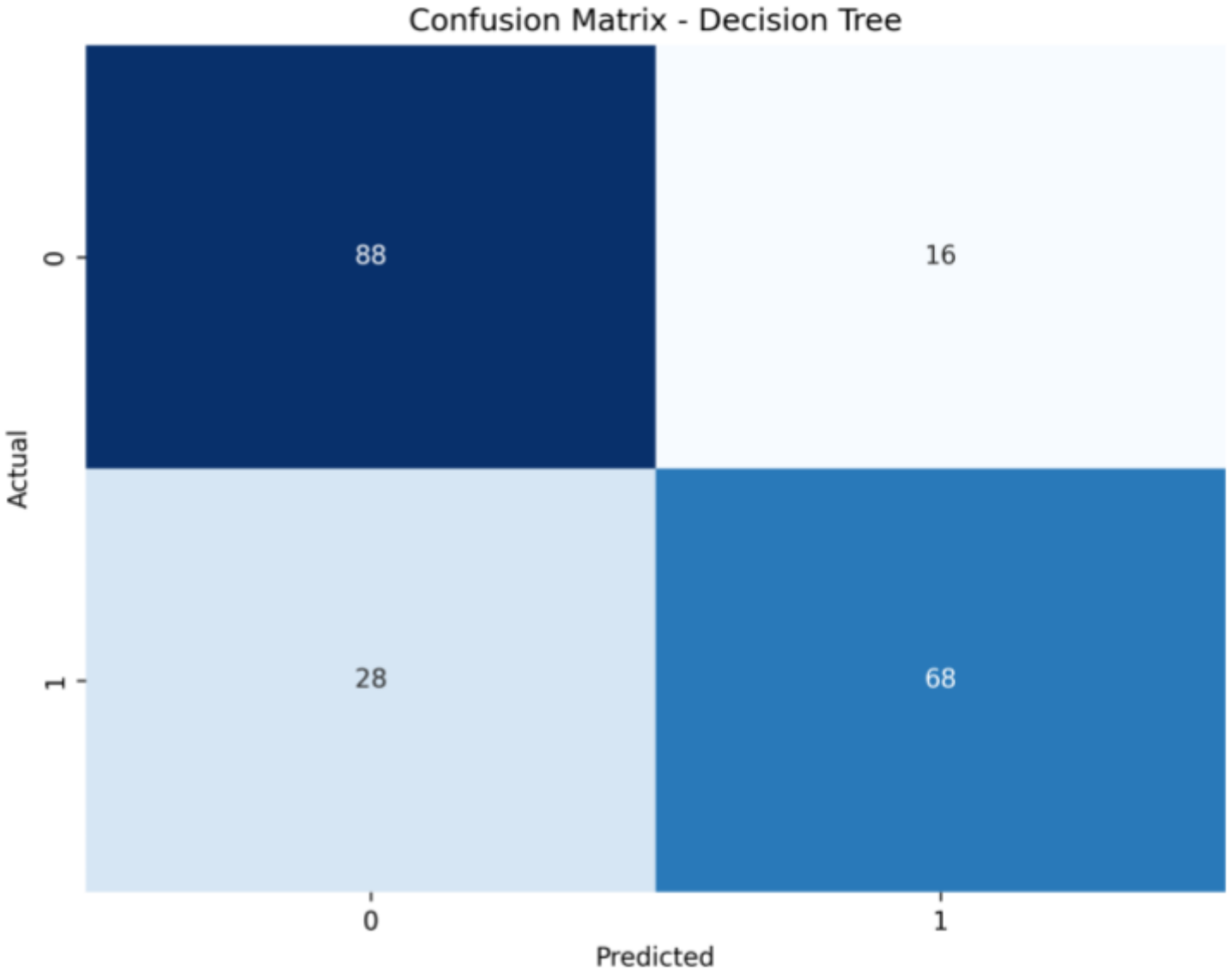

Figure 7: Decision Tree Confusion Matrix

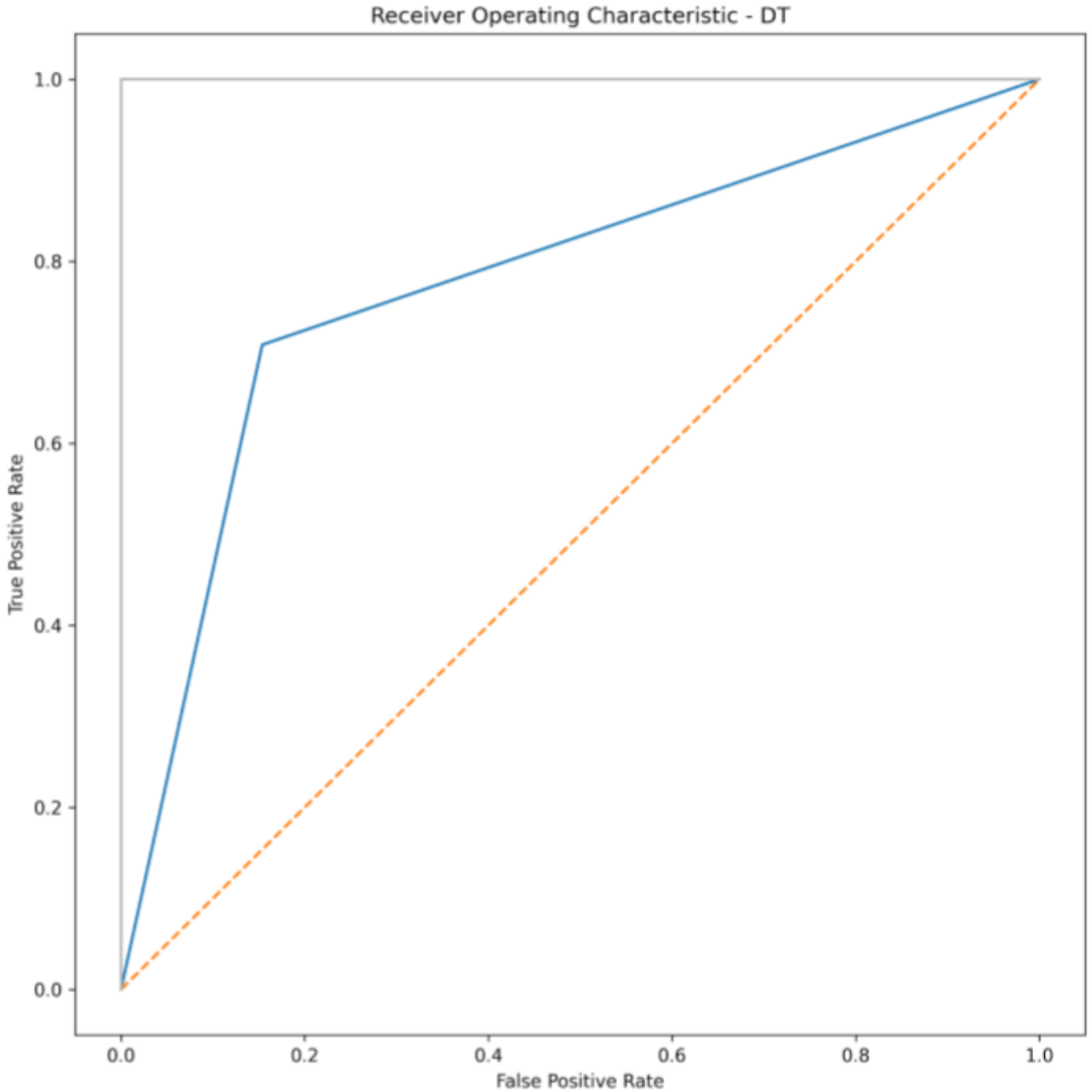

Figure 8: Decision Tree ROC

Naive Bayes

The Naive Bayes model demonstrated excellent performance, achieving an accuracy of 80%. In the positive sentiment class, the model displayed high precision (84%), accurately identifying positive instances. The recall for positive sentiments was 75%, indicating the model's effectiveness in capturing a substantial portion of actual positive sentiments. The F1-score, a balance between precision and recall, reached 79% in the positive class. Similarly, in the negative sentiment class, the model showed robust performance with a precision of 76% and a recall of 84%, resulting in an F1-score of 80%. The overall Receiver Operating Characteristic (ROC) score of 0.796 affirmed the model's ability to discriminate between positive and negative sentiments. Notably, Figures 9 and 10 depict the confusion matrix and ROC curve, respectively. Compared to the Decision Tree model, Naive Bayes exhibited slightly superior overall performance, boasting higher accuracy, recall, F1, and ROC scores.

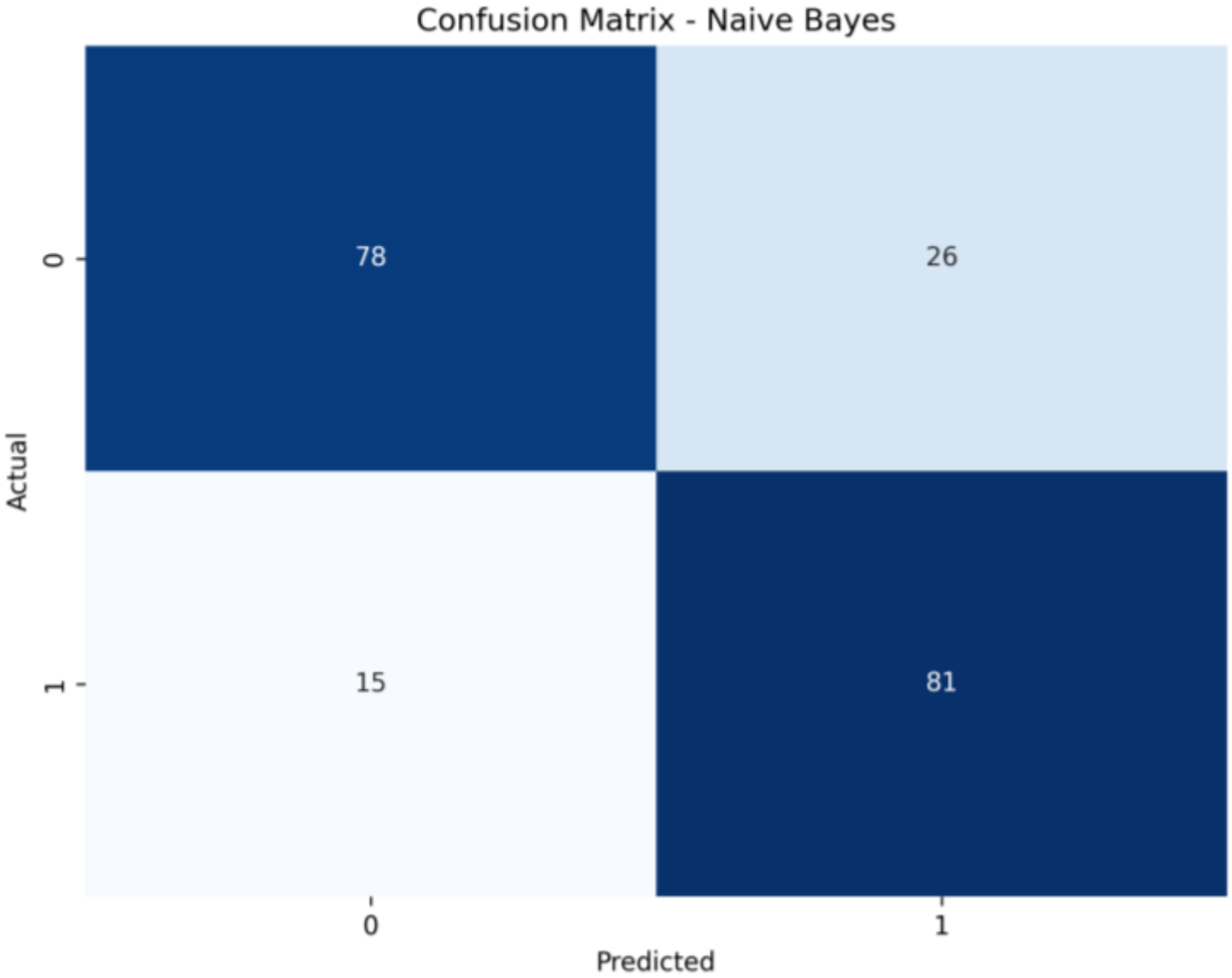

Figure 9: Naive Bayes Confusion Matrix

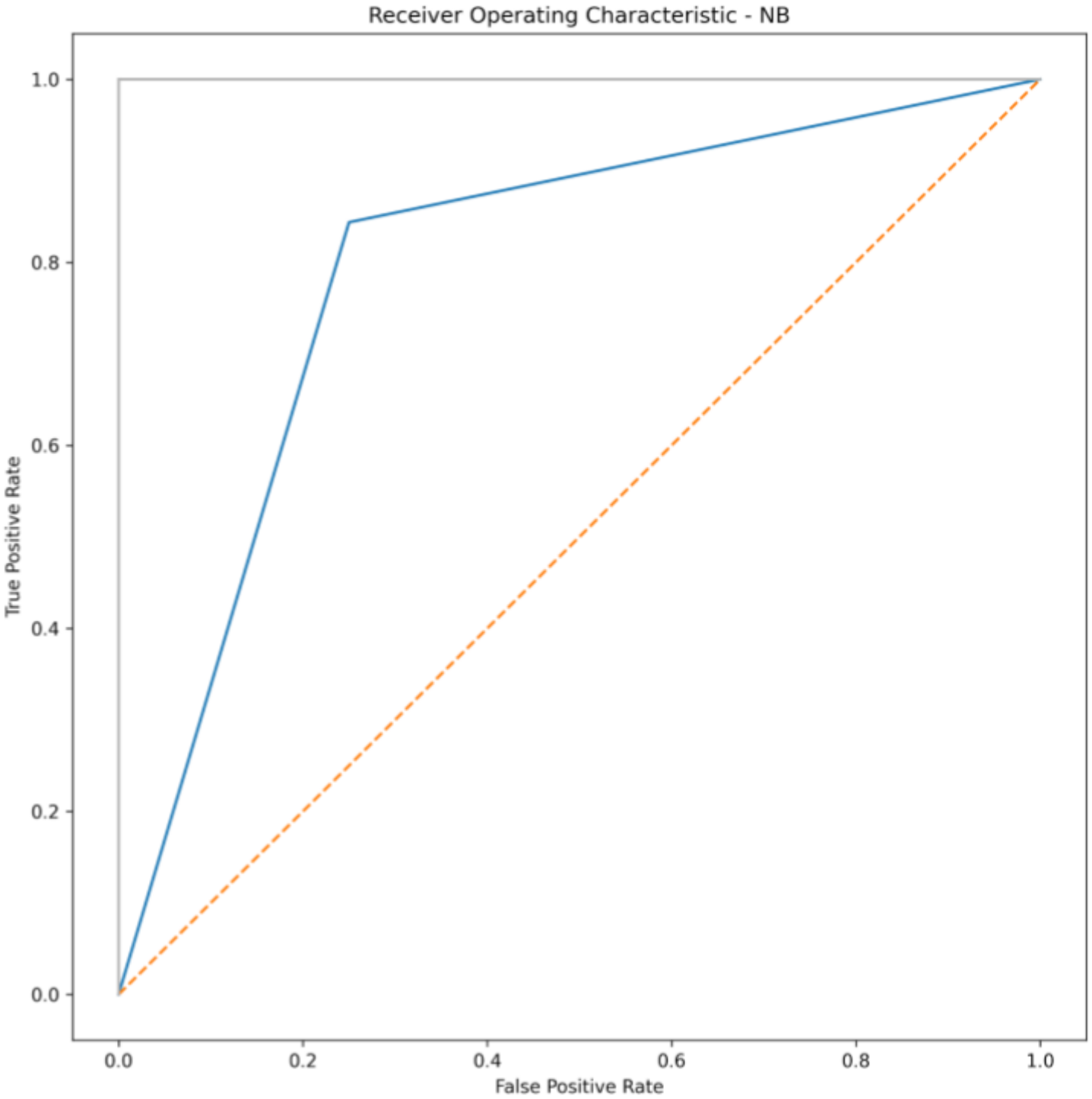

Figure 10: Naive Bayes Roc

SVM

The Support Vector Machine (SVM) model demonstrated solid performance, achieving an accuracy of 79%. In the positive sentiment class, the model exhibited a precision of 78%, signifying its ability to accurately classify positive instances. The recall for positive sentiments was 82%, indicating the model effectively captured a significant portion of actual positive sentiments. The F1-score, a harmonic mean of precision and recall, reached 80% in the positive class. For the negative sentiment class, the SVM model displayed a precision of 79% and a recall of 75%, resulting in an F1-score of 77%. Overall, SVM showcased a balanced performance between precision and recall. When comparing the SVM model with the previously evaluated Decision Tree and Naive Bayes models, SVM falls in between, providing a good trade-off between precision and recall.Figures 11 and 12 provide a visual representation of the performance of this model.

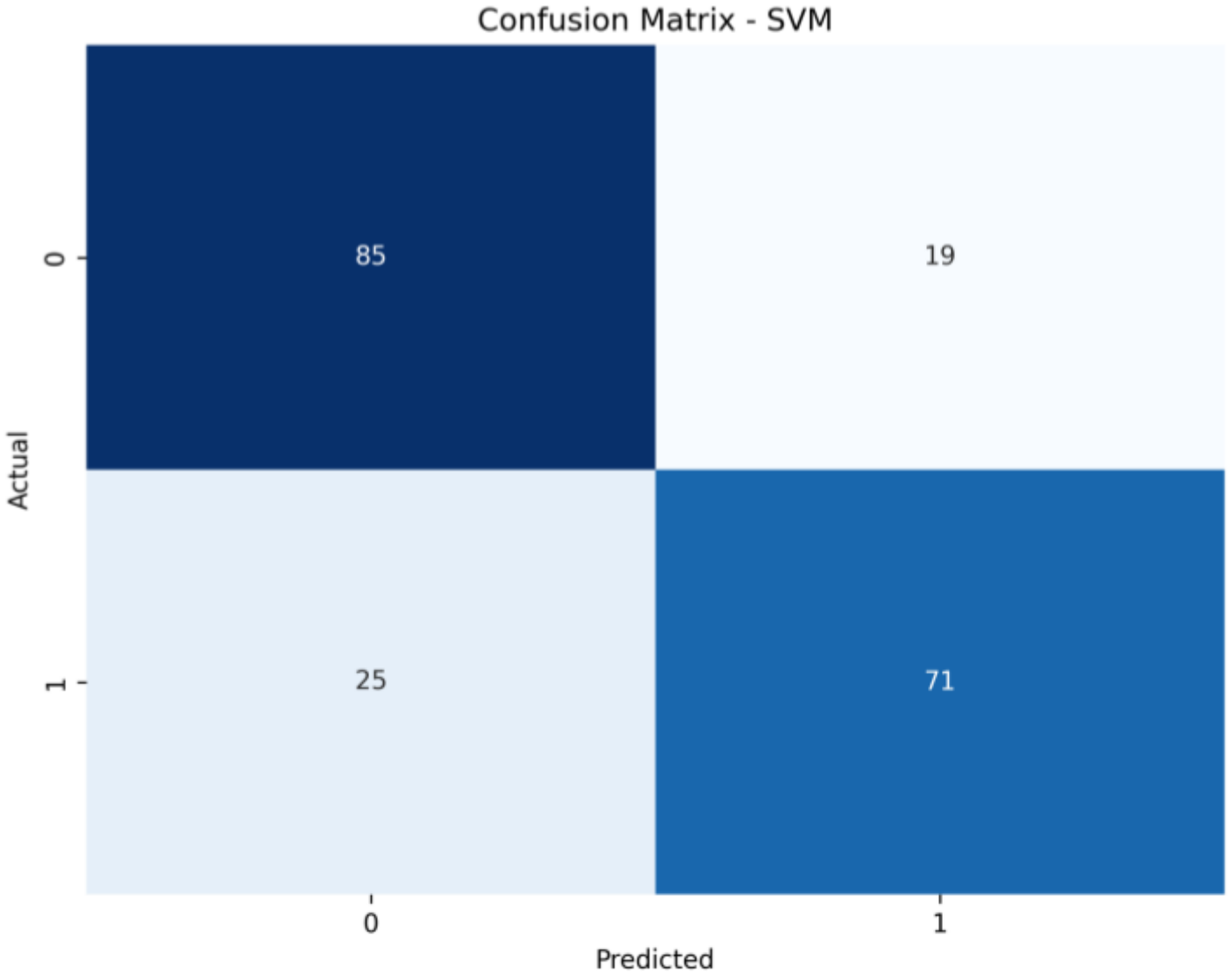

Figure 11: SVM Confusion Matrix

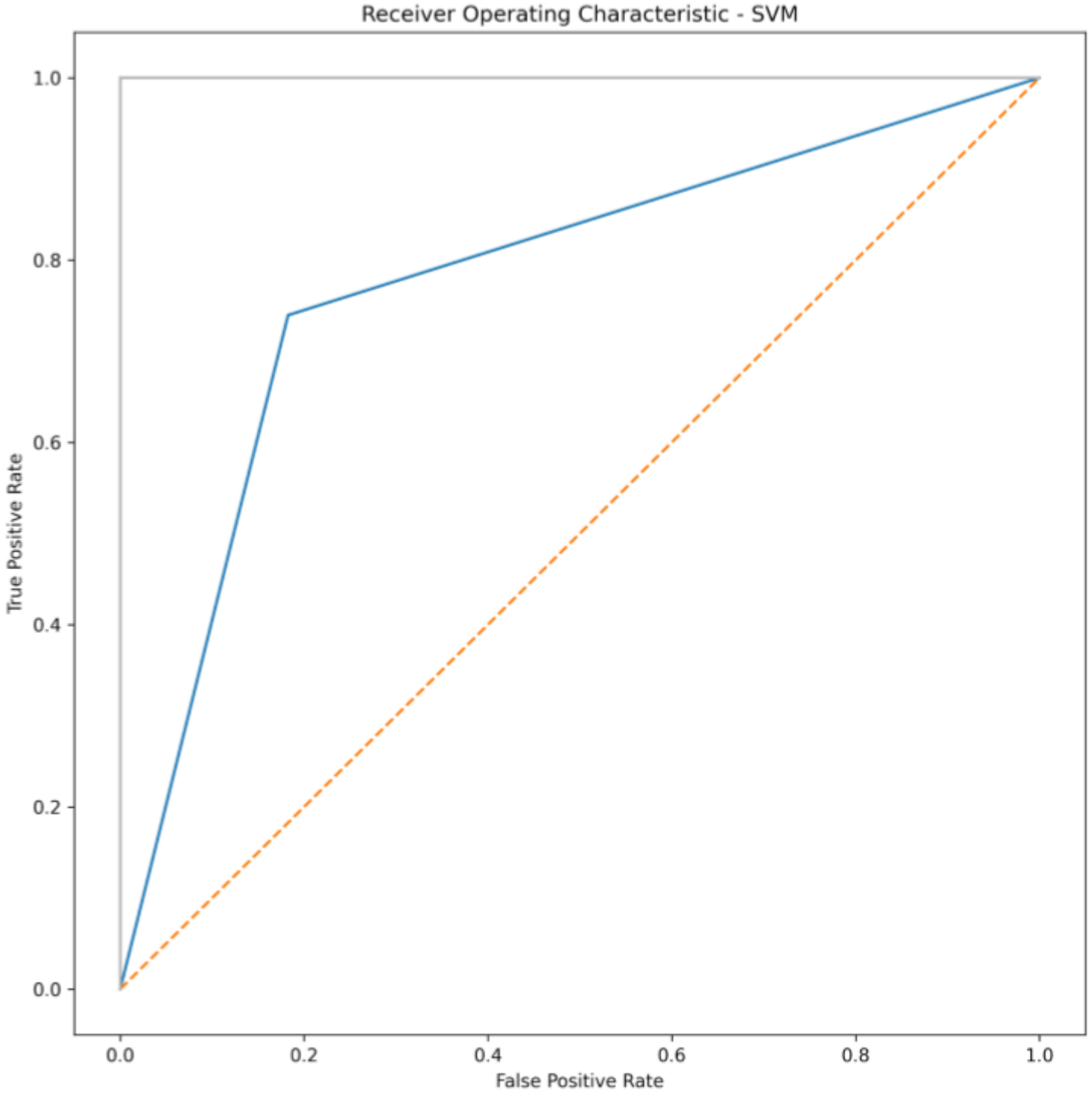

Figure 12: SVM Roc

Random Forest

The Random Forest model delivered robust sentiment classification with an 80% accuracy. In the positive class, it achieved 79% precision and 85% recall, resulting in an 81% F1-score. For the negative class, the model displayed 82% precision, 75% recall, and a 78% F1-score. The overall ROC score was 0.798, showcasing effective discrimination between positive and negative sentiments. Compared to the Decision Tree, Naive Bayes, and SVM models, the Random Forest outperformed all. Figure 13 shows the confusion matrix, and Figure 14 presents the ROC curve for detailed performance analysis.

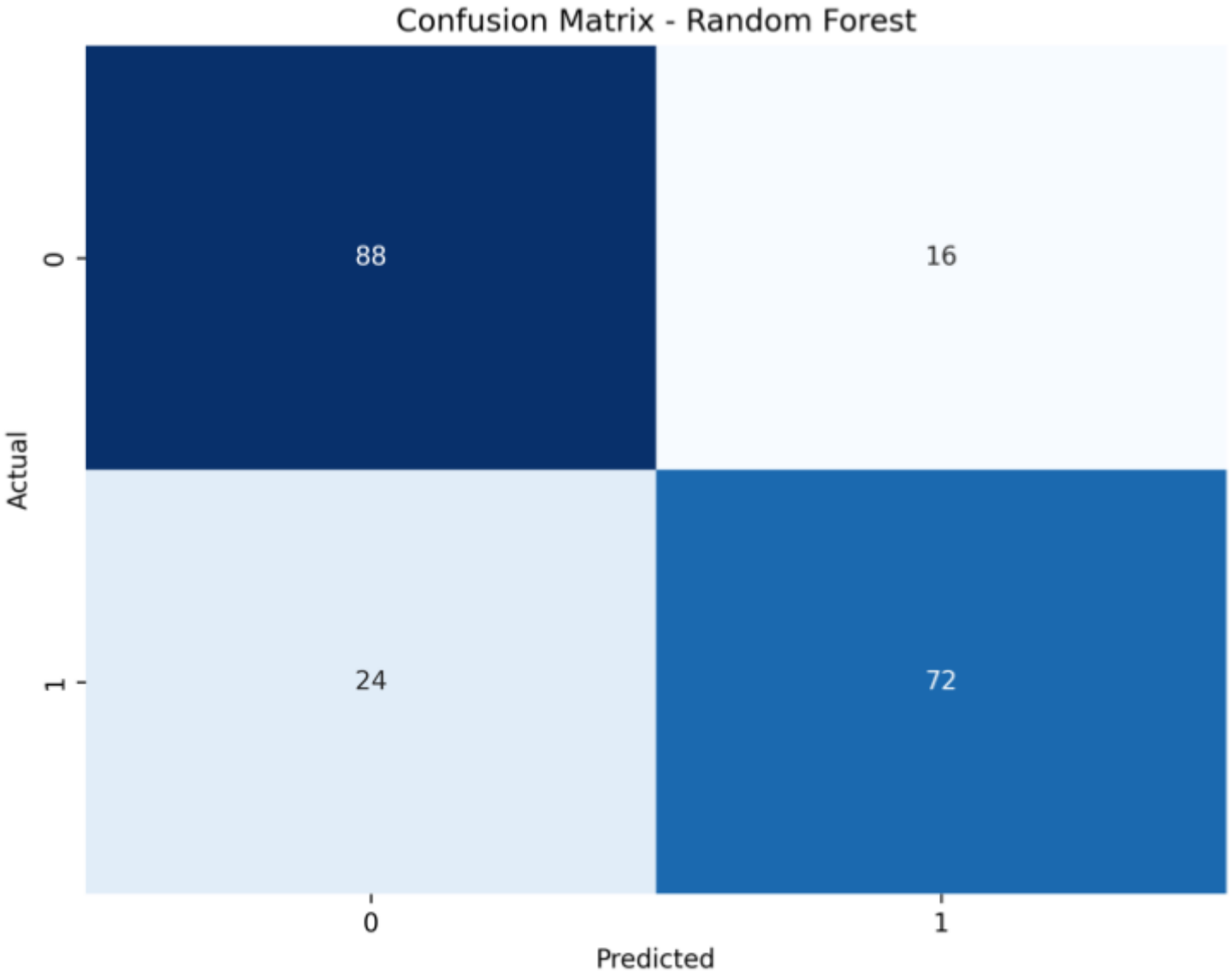

Figure 13: Random Forest Confusion Matrix

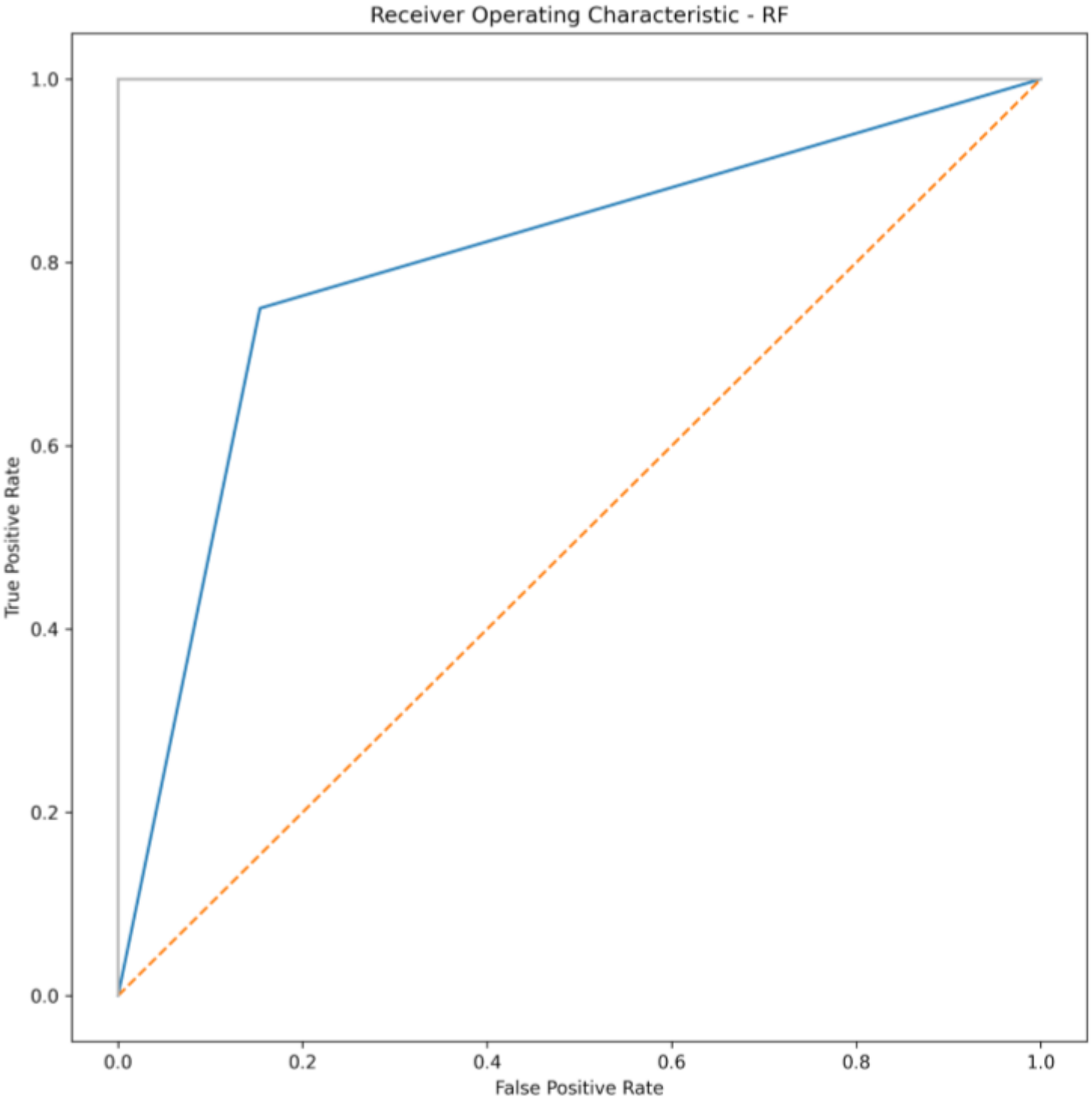

Figure 14: Random Forest Roc

In Figure 15 below, a comparative analysis of the models' accuracy is presented, revealing that all models performed exceptionally closely. Notably, the Random Forest model emerged as the top performer, attaining the highest accuracy among the evaluated models.

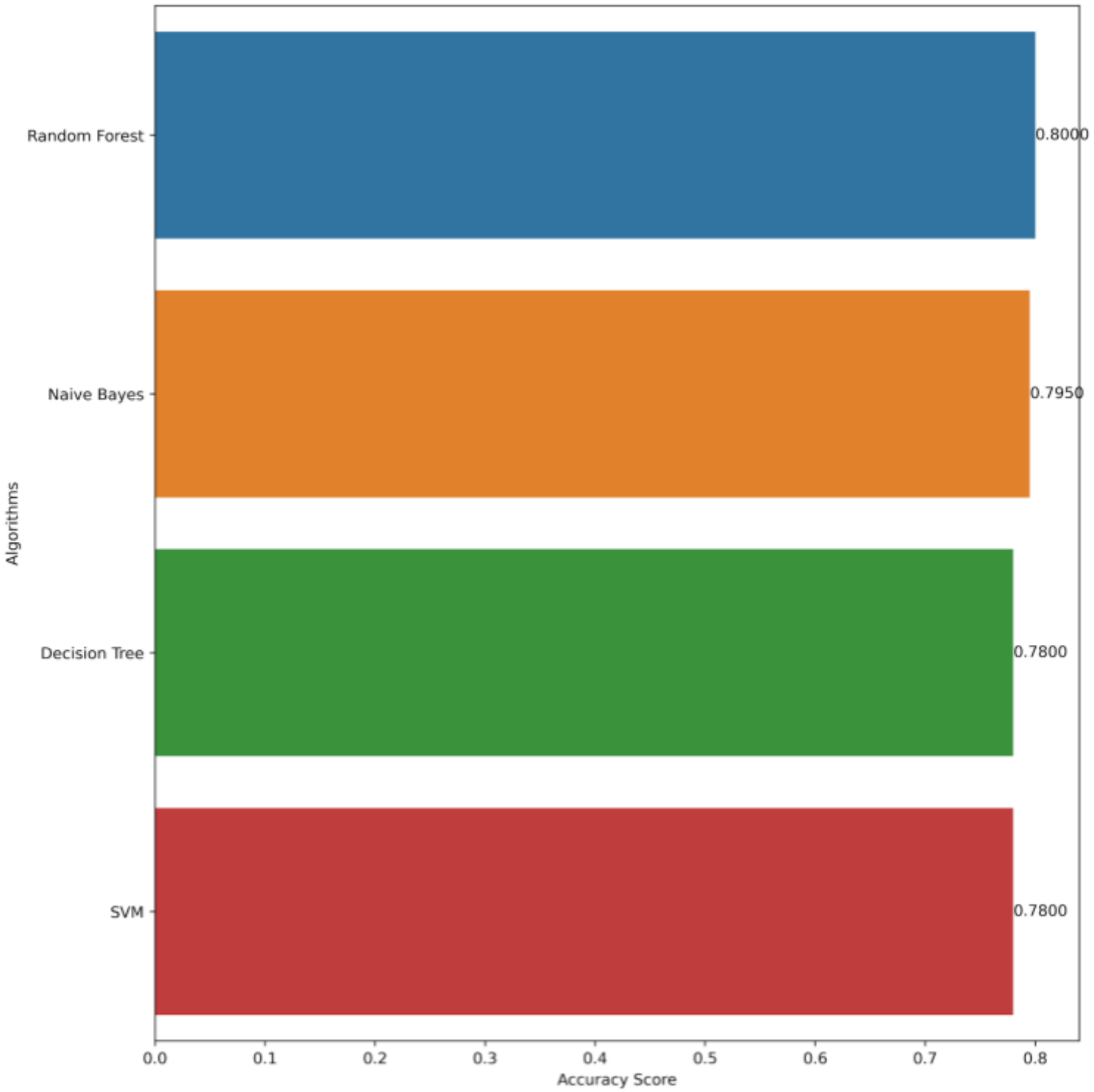

Figure 15: Accuracy Comparison

Part 3: Bert Base Model With Fine-Tuning

The BERT-based model with fine-tuning emerged as the top performer, outclassing all other four models—Random Forest, Decision Tree, Naive Bayes, and SVM. This superior performance is clearly illustrated in Figure 16 and Figure 17, where the table and accuracy comparison plot show that the BERT model dominated on all of the metrics . The exceptional performance of the BERT model can be attributed to its advanced architecture and the fine-tuning process. The model was trained over an increased number of epochs(30), enhancing its ability to capture intricate patterns in the data. Figure 18 provides insights into the training and validation process, demonstrating a consistent decrease in both training and validation loss after each epoch. Simultaneously, accuracy steadily increases, indicating effective learning and generalisation. The success of the BERT-based model underscores its capability to leverage contextual embeddings and adapt to the intricacies of sentiment analysis tasks.

Algorithms	Precision	Recall	F1 Score	Accuracy Score	Auc Score
Bert	0.8617	0.81	0.8351	0.84	0.84
Random Forest	0.8182	0.75	0.7826	0.8	0.7981
Naive Bayes	0.757	0.8438	0.798	0.795	0.7969
Decision Tree	0.8095	0.7083	0.7556	0.78	0.7772
SVM	0.7889	0.7396	0.7634	0.78	0.7784

Figure 16: Bert Comparison

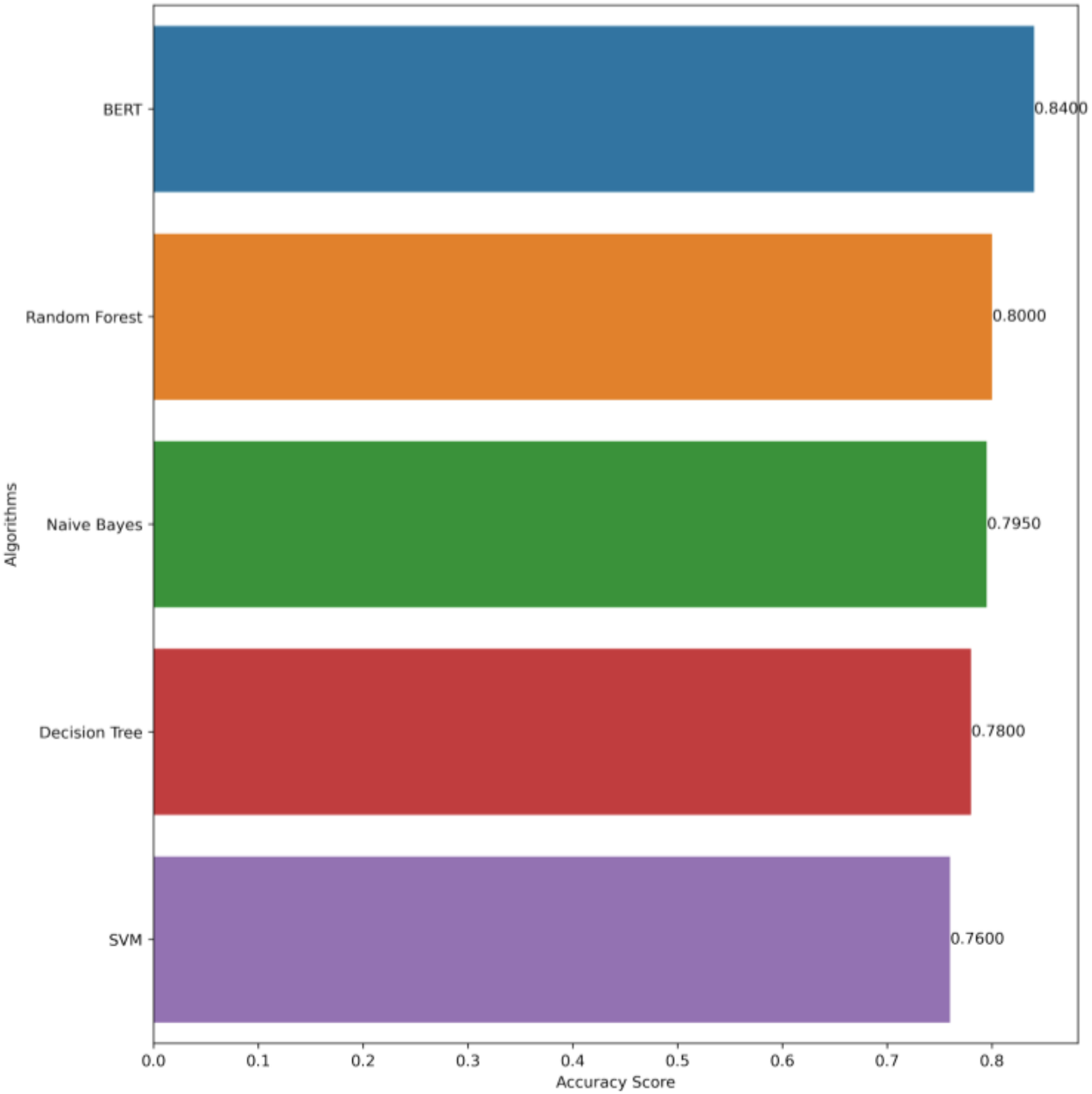

Figure 17: Bert accuracy comparison

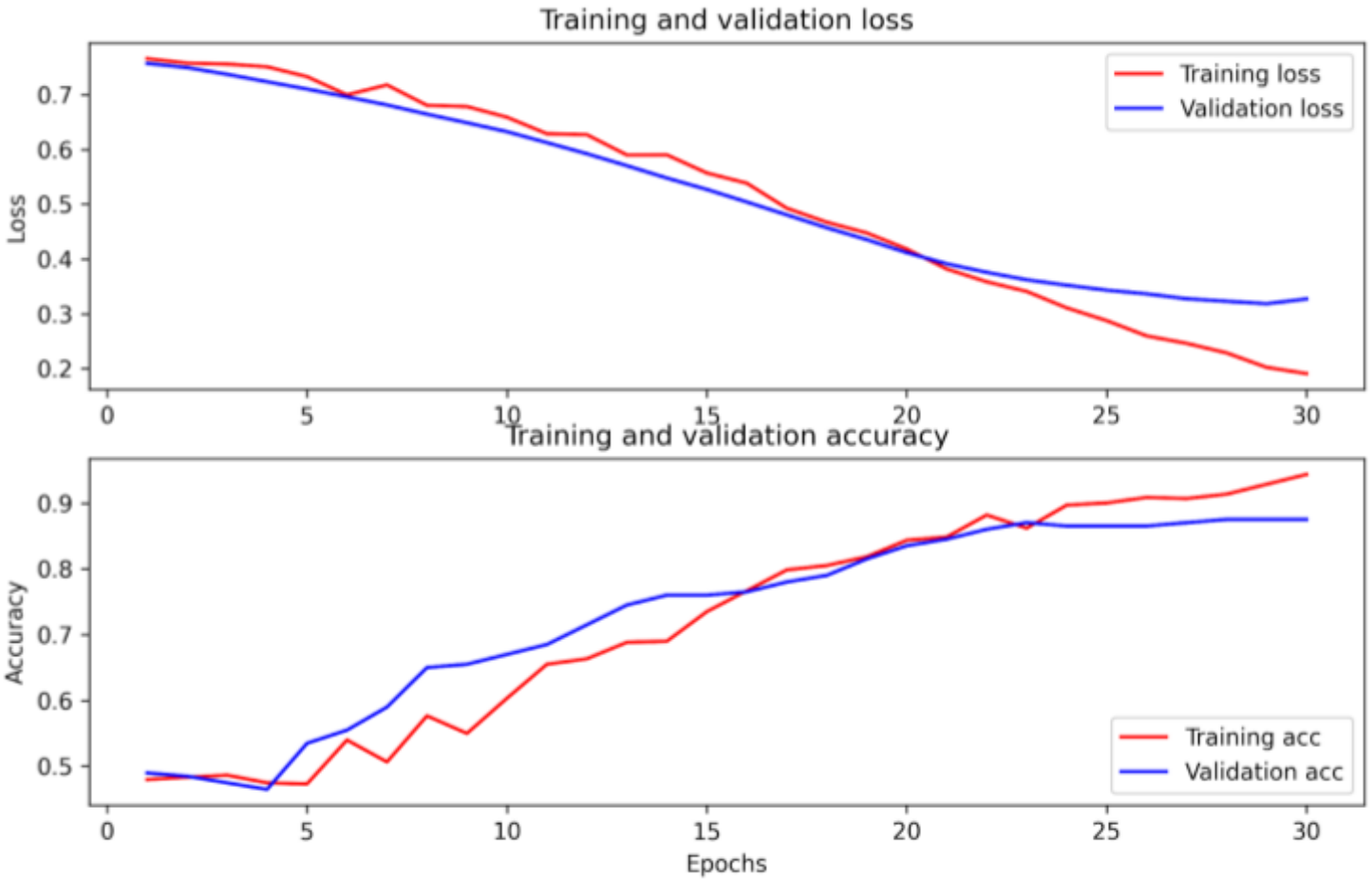

Figure 19: Bert Training and Validation Accuracy and Loss

Part 4: Topic Detection

Lda Model

After topic detection was applied to the data figure 20 shows the Intertopic Distance Map of the five topics detected.

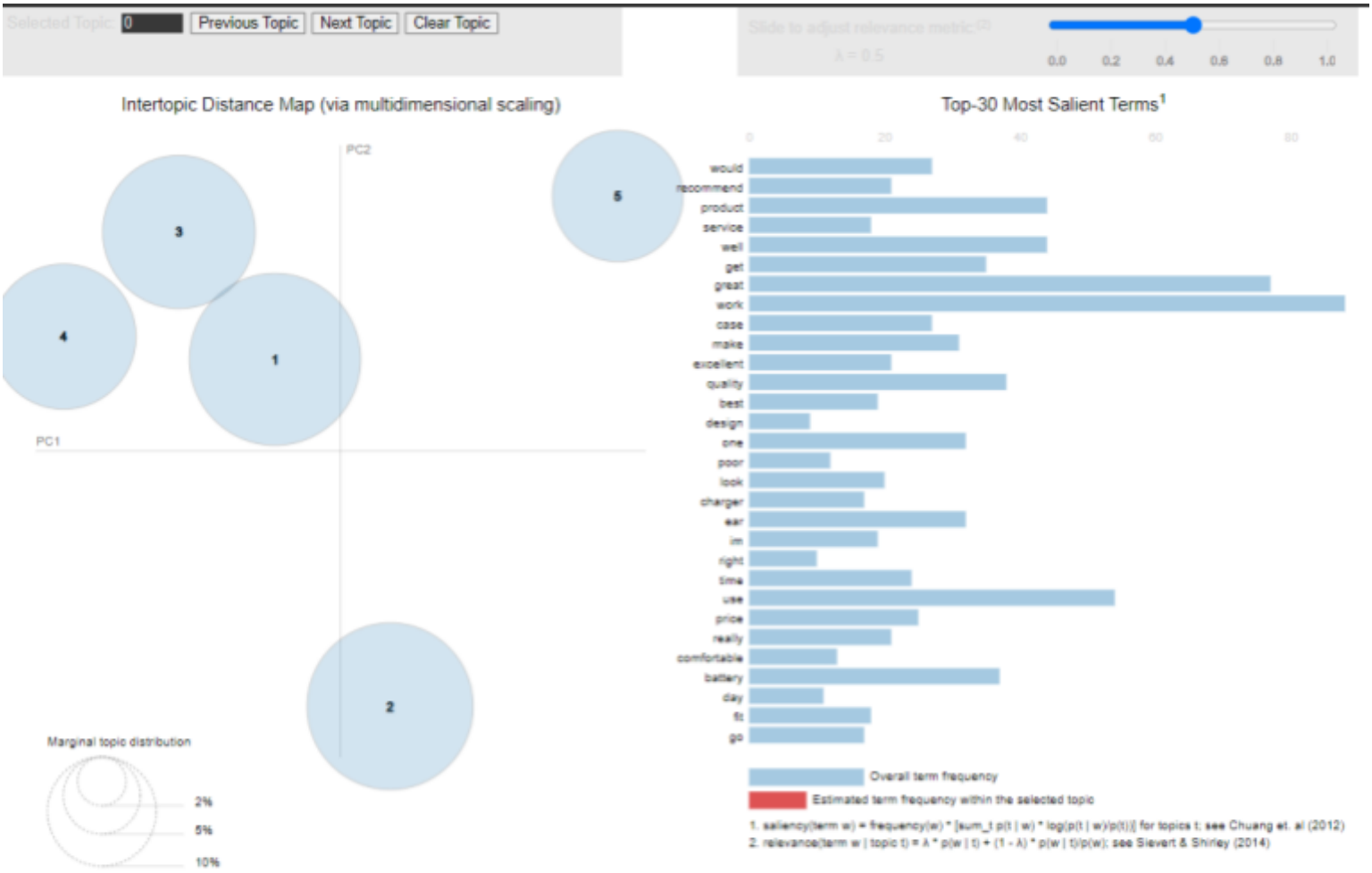

Figure 20: Topic Detection LDA Model

Setting the lambda(λ) parameter to 0.5 in the intertopic distance map strikes a balance between highlighting distinctive terms in individual topics and incorporating common terms shared across multiple topics. This choice ensures that the identified topics are both coherent and interpretable, capturing unique and common aspects of the Amazon reviews dataset.

Topic 1:

Topic 1 centres on the functionality and usability of products, particularly phones, headsets, and related devices. Common terms like "work," "product," and "great" indicate positive sentiments about performance and quality. Expressions like "problem," "waste," and "disappoint" hint at potential concerns, while words like "buy," "money," and "time" suggest discussions on purchasing decisions and overall product value.

Topic 2:
Topic 2 captures a mix of positive and negative sentiments, highlighting user experiences with phones and headsets. The topic quality appears moderate, offering insights into user preferences and potential pain points associated with these products.

Topic 3:
Topic 3 revolves around various aspects, including battery life, case quality, and price considerations. It also touches on user experiences with phones, focusing on attributes like "great" and "comfortable." The discussion spans different scenarios, such as car usage and charging, indicating a diverse range of considerations within this topic.

Topic 4:
Topic 4 explores design, functionality, and user experience, with terms like "well," "good," and "comfortable" reflecting positive sentiments. It also touches on potential flaws and the need for replacement, indicating a mix of positive and critical viewpoints. The topic covers diverse aspects, including phone features, company considerations, and customer contact. Despite a somewhat diverse vocabulary, the overall quality is moderate, encompassing sentiments about design, functionality, and user satisfaction.

Topic 5:

Topic 5 focuses on user recommendations and opinions, featuring positive terms like "recommend," "would," and "love," as well as negative expressions such as "poor," "horrible," and "crap." Encompassing product quality, audio experience, customer service, and phone features, the topic's inclusion of both positive and negative terms enhances its comprehensiveness, capturing diverse user opinions. The quality of this topic is noteworthy, showcasing a range of sentiments and recommendations expressed by users.

Bert Topic Detection(3)

BERT-based topic detection involves using the Bidirectional Encoder Representations from Transformers (BERT) model, fine-tuned on a specific dataset, to identify and classify topics within a collection of text, revealing the underlying themes present in the data. Figure 21 shows each topic detected in the dataset by Bertopic model.

Figure 21: Bert Topic Detection

Topic 0:
This topic appears to revolve around positive sentiments, mentioning terms like "great," "I've," and "work." It suggests a generally favourable sentiment, but the specific context may vary.

Topic 1:
This topic relates to comfort and service, with terms like "fit," "nice," and "comfortable." It seems to highlight aspects related to user experience and satisfaction.

Topic 2:
Focused on audio quality, this topic mentions terms like "quality voice," "poor," and "noise." It addresses the audio performance of the mentioned product or service.

Topic 3:
The terms in this topic, such as "great," "doesn't," "fine," and "perfectly," suggest a positive sentiment, possibly indicating satisfaction with specific features or aspects.

Topic 4:
Centred around Bluetooth usage and calls, terms like "Bluetooth," "use," and "call" indicate a focus on connectivity and communication aspects.

Topic 5:
Expressing disappointment and problems, this topic includes terms like "disappointment," "order," and "problem." It may represent negative sentiments or issues experienced by users.

Topic 6:
Pertaining to purchasing decisions, terms like "purchase," "buying," and "buyer" suggest discussions around the buying process and considerations for potential customers.

Topic 7:
This topic touches on the longevity of a product or service, with terms like "life," "long," and "die." It might involve discussions about the durability or lifespan of the mentioned item.

References

1. Sklearn.ensemble.randomforestclassifier. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.https://scikit-learn.org/stable/about.html#citing-scikit-learn
3. Grootendorst, M. P. Bertopic - Bertopic. BERTopic - BERTopic. https://maartengr.github.io/BERTopic/api/bertopic.html#bertopic._bertopic.BERTopic.fit