Intelligent Data and Text Analytics Coursework 2

Text Dataset Analysis

Submitted by
2293815

Submitted to:
University of Portsmouth

Date of Submission:
January 20, 2025

Module: Intelligent Data and Text Analytics

Contents

List of Tables

List of Figures

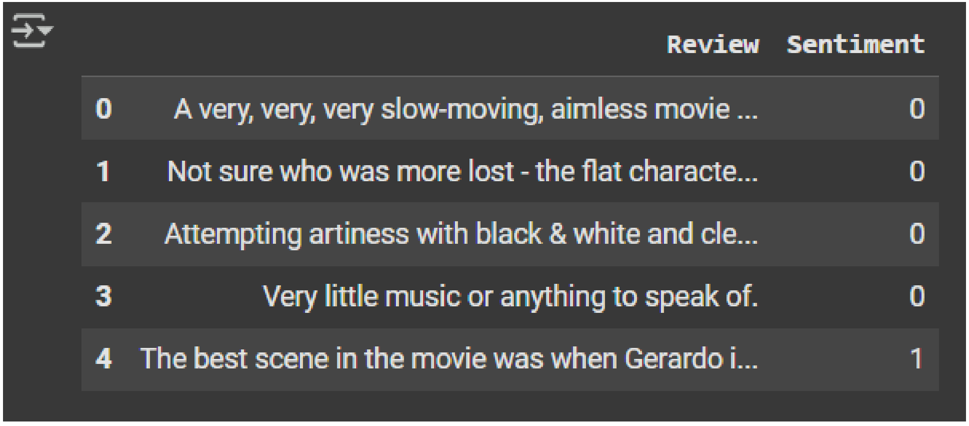
Chapter 1

Introduction

1.1 Dataset Description

I am using the imdb-labelled dataset for this coursework. The dataset has 748 sentences in total. The dataset looks like Figure1.1. The Sentiment column has 2 values, 0 or 1, which means negative or positive sentiment of the comment. My goal in this report is to apply different techniques and algorithms to this dataset and analyze the results.

Figure 1.1: Dataset Overview



	Review	Sentiment
0	A very, very, very slow-moving, aimless movie ...	0
1	Not sure who was more lost - the flat characte...	0
2	Attempting artiness with black & white and cle...	0
3	Very little music or anything to speak of.	0
4	The best scene in the movie was when Gerardo i...	1

Chapter 2

Preprocessing Textual Data

After uploading the .txt file to google colab, I transformed the .txt file to .csv file. Now, I can apply traditional Text preprocessing tasks. This will make the dataset more cleaned and suitable for the Machine Learning models to process it. The more the data will be cleaned, the more better results we will get. Mohasseb, 2025

2.1 Removing Punctuation

Shown in Table2.1, the punctuation marks like comma, semicolon, etc are removed.

Before Preprocessing	After Preprocessing
A very,very, very slow-moving, aimless movie about a distressed, drifting young man.	a very very very slowmoving aimless movie about a distressed drifting young man
Very little music or anything to speak of.	very little music or anything to speak of
Plus, it was well-paced and suited its relatively short run time.	plus it was wellpaced and suited its relatively short run time

Table 2.1: Before and After Preprocessing (Removing Punctuation)

2.2 Removing Numbers

In this stage, any numerical characters are removed, as shown in Table2.2.

Before Preprocessing	After Preprocessing
This if the first movie I've given a 10 to in years.	this if the first movie ive given a to in years
I gave it a 10	i gave it a
IMDB ratings only go as low 1 for awful	imdb ratings only go as low for awful

Table 2.2: Before and After Preprocessing (Removing Numbers)

2.3 Removing Stop Words

In English language the stop words are "and," "the," "is," "in," etc, and in this step, I filtered these out as shown in Table2.3. OpenAI, 2025

Before Preprocessing	**After Preprocessing**
a very very very slowmoving aimless movie	slowmoving aimless movie
not sure who was more lost	sure lost
the best scene in the movie was	best scene movie

Table 2.3: Before and After Preprocessing (Stop Word Removal)

2.4 Changing Text Case

All capital letters are converted into small letters in this step as shown in Table2.4.

Before Preprocessing	**After Preprocessing**
A very, very, very slow-moving, aimless movie	a very very very slowmoving aimless movie
Not sure who was more lost	not sure who was more lost
Attempting artiness with black	attempting artiness with black

Table 2.4: Before and After Preprocessing (Case Lowering)

2.5 Lemmatization

In this step, all the words are converted to their base form, like shown in Table2.5.

Before Preprocessing	**After Preprocessing**
best scene movie gerardo trying find song	best scene movie gerardo try find song
loved casting jimmy buffet	love cast jimmy buffet
attempting artiness black white	attempt artiness black white

Table 2.5: Before and After Preprocessing (Lemmatization)

2.6 Special Character Removal (EXTRA)

In this step, I removed any special characters occuring in the reviews like shown in Table2.6. Developers, 2025

Before Preprocessing	**After Preprocessing**
It's practically perfect in all of them Â	practically perfect
The script is Â	script
I'll even say it again Â– this is torture	ill even say torture

Table 2.6: Before and After Preprocessing (Special Character Removal)

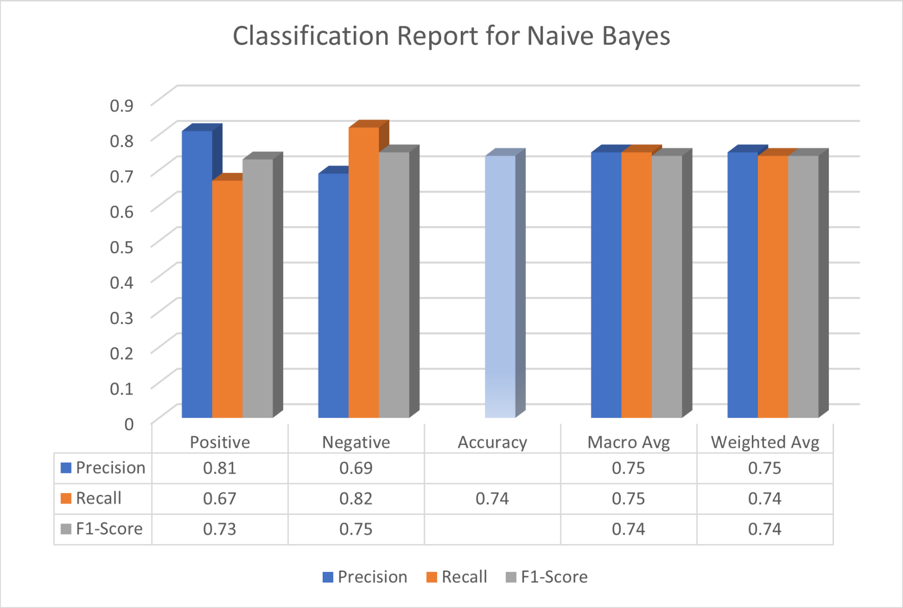
Chapter 3

Classification Using Bag-of-Words

By using Bag of Words, I transformed every sentence into vectors or matix for the classification models to understand. Then we split the dataset, 500 were for training, and remaining 248 were for testing part. OpenAI, 2025 Mohasseb, 2025

3.1 Naive Bayes

The Naive Bayes model achieved an accuracy of 74.19% . It demonstrates a good balance between precision and recall for both the positive and negative classes. The positive class has a slightly higher precision (0.81), indicating fewer false positives, but its recall is lower (0.67), meaning some positive instances are missed. Conversely, the negative class has a lower precision (0.69) but a higher recall (0.82), showing it identifies most negative instances correctly. The F1-scores for both classes are close (0.73 for positive, 0.75 for negative), suggesting the model performs fairly well across both categories. The results are shown in Figure3.1. Mohasseb, 2025

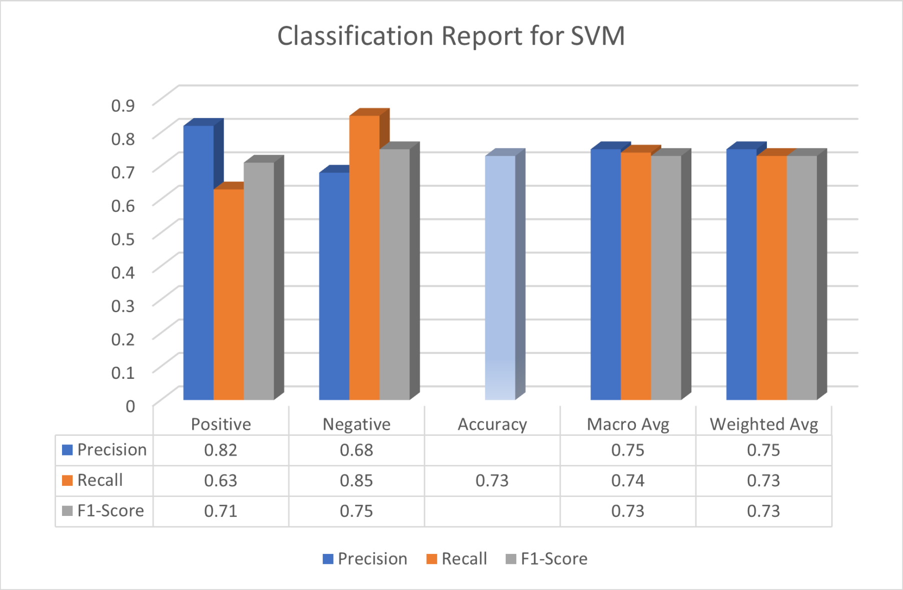


	Positive	Negative	Accuracy	Macro Avg	Weighted Avg
Precision	0.81	0.69		0.75	0.75
Recall	0.67	0.82	0.74	0.75	0.74
F1-Score	0.73	0.75		0.74	0.74

Figure 3.1: Classification Report for Naive Bayes

3.2 Support Vector Machine (SVM)

The SVM achieved an accuracy of 0.73 on the classification task. The model performed relatively better on predicting the Negative class with a high recall (0.85) and a balanced F1-score (0.75). The Positive class had a lower recall (0.63), but a higher precision of 0.82. Overall, the model showed a good balance between the two classes in terms of precision and recall, with macro and weighted average F1-scores around 0.73. The classification report is shown in Figure3.2. Mohasseb, 2025

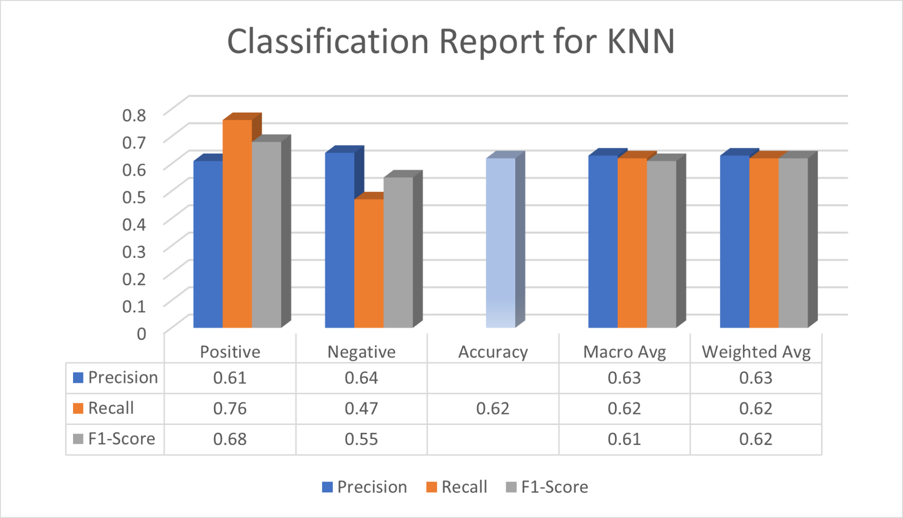


	Positive	Negative	Accuracy	Macro Avg	Weighted Avg
■ Precision	0.82	0.68		0.75	0.75
■ Recall	0.63	0.85	0.73	0.74	0.73
■ F1-Score	0.71	0.75		0.73	0.73

Figure 3.2: Classification Report for SVM

3.3 K Nearest Neighbors

The KNN model achieved an accuracy of 0.625, which is lower compared to the SVM model. The model performed relatively better on predicting the Positive class with a higher recall of 0.76, but the precision is lower (0.61). The Negative class showed a lower recall (0.47), leading to a lower F1-score (0.55). The overall performance metrics, including macro and weighted averages, indicate that the model struggles with balancing the two classes effectively. The classification report is shown in Figure3.3. Mohasseb, 2025

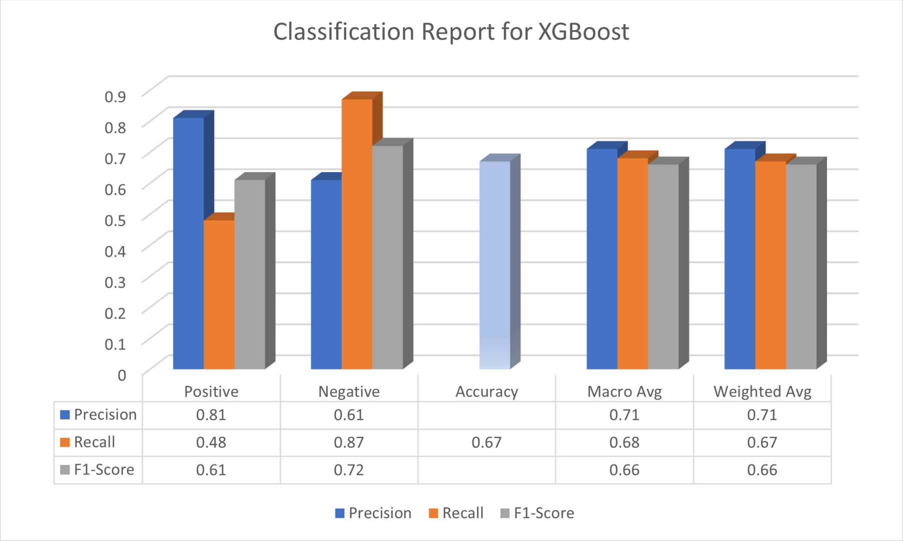


	Positive	Negative	Accuracy	Macro Avg	Weighted Avg
Precision	0.61	0.64		0.63	0.63
Recall	0.76	0.47	0.62	0.62	0.62
F1-Score	0.68	0.55		0.61	0.62

Figure 3.3: Classification Report for KNN

3.4 XGBoost (EXTRA)

The XGBoost model achieved an accuracy of 0.67. The model performed better in predicting the Negative class, with high recall (0.87) and a decent F1-score (0.72). However, the Positive class had a lower recall (0.48), which resulted in a reduced F1-score (0.61). Although the model showed strong precision for the Positive class (0.81), it struggled with recall, leading to a lower overall performance compared to other models. The classification report is shown in Figure3.4. Developers, 2025



	Positive	Negative	Accuracy	Macro Avg	Weighted Avg
Precision	0.81	0.61		0.71	0.71
Recall	0.48	0.87	0.67	0.68	0.67
F1-Score	0.61	0.72		0.66	0.66

Figure 3.4: Classification Report for XGBoost

3.5 Comparison of Algorithms

The Naive Bayes (NB) model with BoW achieved the highest accuracy (0.74) and also demonstrated a balanced F1-score of 0.74 for both classes, making it the most well-rounded model. The SVM model with BoW followed with an accuracy of 0.73, but its

F1-score (0.73) was slightly lower than NB's. XGBoost, with an accuracy of 0.67, showed a significant drop in recall for the Positive class (0.48), resulting in a lower F1-score (0.66), while KNN performed the worst with an accuracy of 0.625 and a noticeably low F1-score (0.61). Overall, Naive Bayes and SVM performed the best in terms of balancing both classes, as indicated by their higher F1-scores. The comparison of models is shown in Figure3.5. OpenAI, 2025

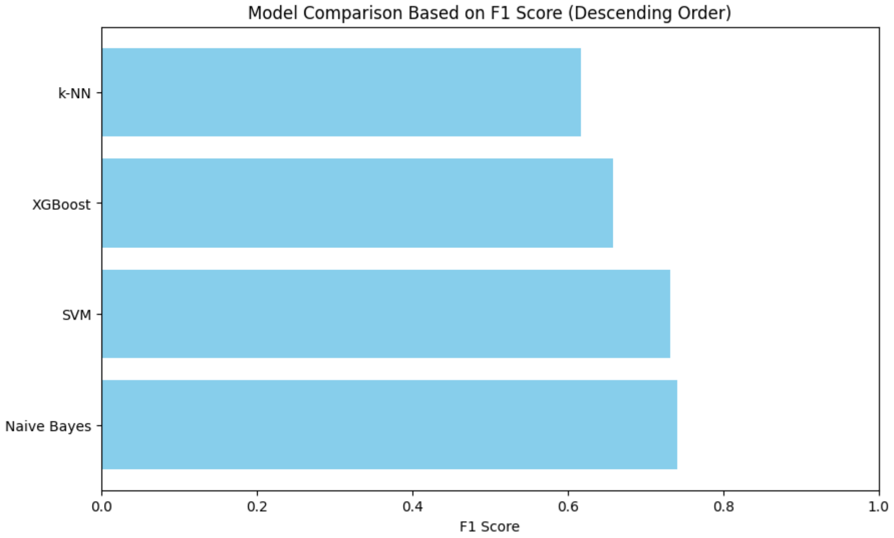


Figure 3.5: Comparison of the Classification Models

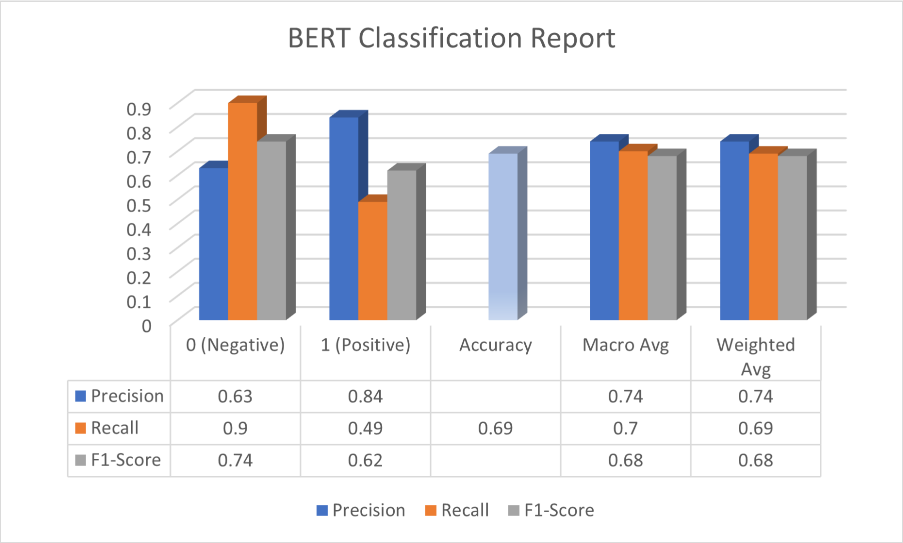
I considered F1-score, because it combines both precision and recall into a single metric, offering a balanced view of the model's performance. Mohasseb, 2025

Chapter 4

BERT Classifiers

4.1 Classification using BERT

The BERT model achieved an accuracy of 69.3% and an AUC of 0.70, indicating moderate performance. It performed better in identifying Class 0 (recall: 0.90, F1-score: 0.74) compared to Class 1 (recall: 0.49, F1-score: 0.62), suggesting difficulty in correctly predicting Class 1 instances. While precision for Class 1 (0.84) was high, its low recall indicates the model is conservative, likely missing many actual positives. Overall, the macro and weighted F1-scores of 0.68 show balanced but not very optimal performance. The Classification Report is shown in Figure4.1. Mohasseb, 2025

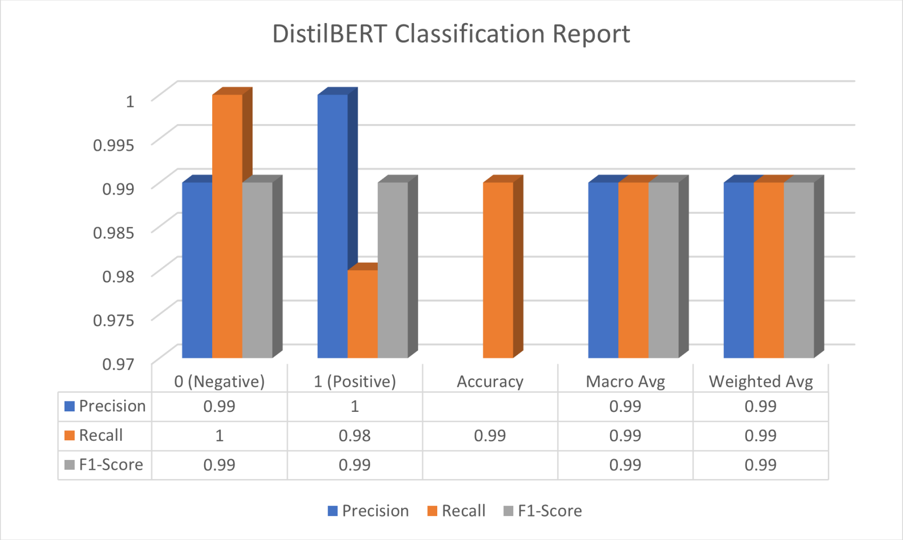


	0 (Negative)	1 (Positive)	Accuracy	Macro Avg	Weighted Avg
🟦 Precision	0.63	0.84		0.74	0.74
🟧 Recall	0.9	0.49	0.69	0.7	0.69
⬜ F1-Score	0.74	0.62		0.68	0.68

Figure 4.1: Classification using BERT

4.2 Classification using DistilBERT (EXTRA)

DistilBERT achieved exceptional performance with an overall accuracy of 99% . Both classes had high precision, recall, and F1-scores (Class 0: 0.99, Class 1: 0.99), indicating that the model performs equally well for identifying true positives and minimizing false positives across both classes. The macro and weighted averages for precision, recall, and F1-score are also 0.99, confirming balanced and robust performance. This result reflects near-perfect classification, suggesting excellent generalization and minimal errors. The classification report for DistilBERT is shown in Figure4.2. OpenAI, 2025 Developers, 2025



	0 (Negative)	1 (Positive)	Accuracy	Macro Avg	Weighted Avg
Precision	0.99	1		0.99	0.99
Recall	1	0.98	0.99	0.99	0.99
F1-Score	0.99	0.99		0.99	0.99

Figure 4.2: Classification using DistilBERT

4.3 Comparison Between the Two BERT-based Models

From Figure4.3 it is seen that BERT accuracy increases steadily, surpassing validation accuracy in later epochs. Validation Accuracy shows consistent improvement, closely tracking training accuracy throughout training. However, it slightly lags behind training accuracy towards the final epochs. OpenAI, 2025

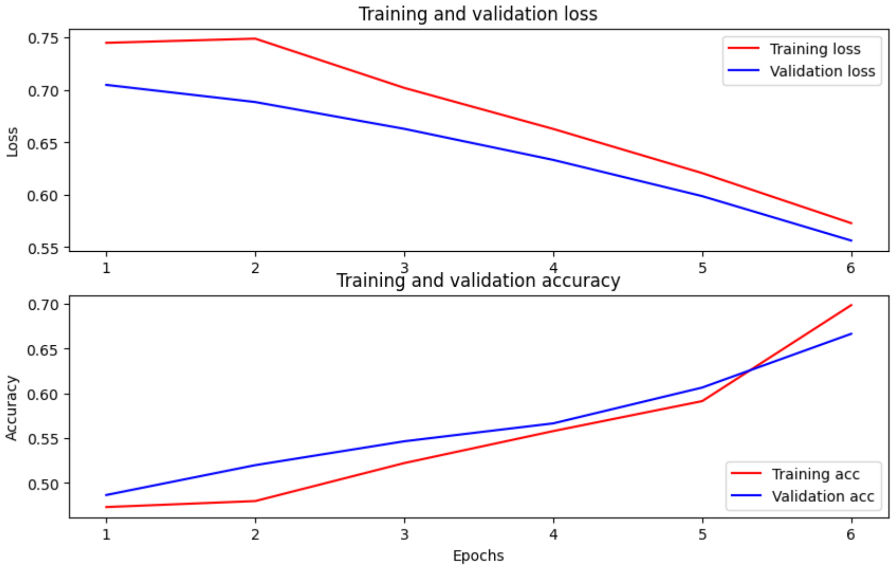


Figure 4.3: Loss Accuracy Curves of BERT

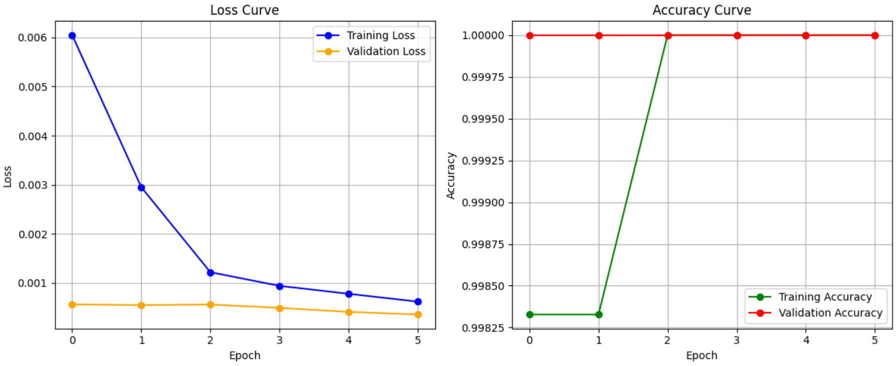


Figure 4.4: Loss Accuracy Curves of DistilBERT

And in Figure4.4, in DistilBERT there is consistent reduction in training loss and increase in accuracy indicate the model was trained successfully. The validation loss and accuracy suggest no significant overfitting or underfitting, with near-perfect performance on both training and validation sets. OpenAI, 2025

In conclusion, DistilBERT is highly efficient, achieving fast accuracy and generalization with fewer epochs.

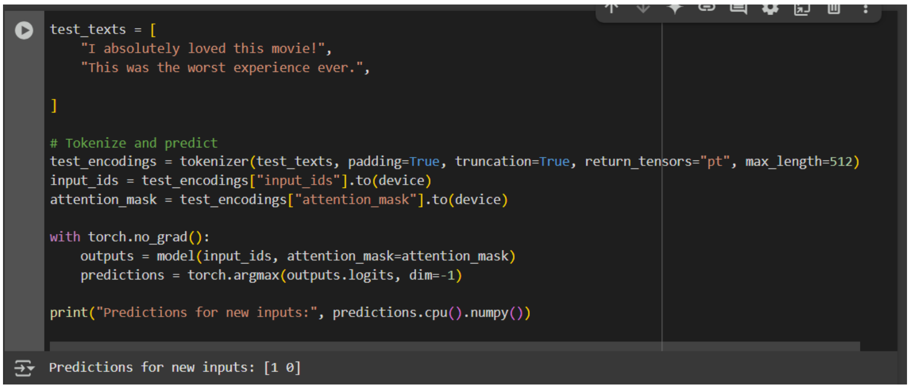
4.4 Results and Comparison with Traditional ML Models

Table4.1 compares all of the model in the basis of their accuracy.

Model	**Accuracy (%)**	**Observations**
Naive Bayes	74.19	Moderate performance, slightly better than SVM, KNN, XGBoost, and BERT.
SVM	73.00	Similar to Naive Bayes, showing competitive accuracy for traditional ML.
KNN	62.50	Lowest performance among the models, likely due to sensitivity to noise and data sparsity.
XGBoost	67.00	Performs better than KNN but falls short compared to Naive Bayes and SVM.
BERT	69.30	Slightly better than XGBoost but underperforms compared to Naive Bayes and SVM.
DistilBERT	99.00	Exceptional performance, far surpassing all other models due to its advanced architecture.

Table 4.1: Model Accuracy Comparison

In conclusion, DistilBERT is the clear winner. I also made predictions with new texts with the DistilBERT model which is shown in Figure4.5, in which the model predicted correctly. Developers, 2025 OpenAI, 2025



```
test_texts = [
    "I absolutely loved this movie!",
    "This was the worst experience ever.",

]

# Tokenize and predict
test_encodings = tokenizer(test_texts, padding=True, truncation=True, return_tensors="pt", max_length=512)
input_ids = test_encodings["input_ids"].to(device)
attention_mask = test_encodings["attention_mask"].to(device)

with torch.no_grad():
    outputs = model(input_ids, attention_mask=attention_mask)
    predictions = torch.argmax(outputs.logits, dim=-1)

print("Predictions for new inputs:", predictions.cpu().numpy())
```

```
Predictions for new inputs: [1 0]
```

Figure 4.5: Prediction using DistilBERT

Chapter 5

Task 4: Topic Detection

In this section, we use topic detection algorithms to group texts and sentences based on specific topics.

5.1 Latent Dirichlet Allocation (LDA)

For Topic Detection, I used 10 as a parameter, so there are 10 topics, from 0 to 9. Below, I am discussing each point.

Topic 0: Disappointing Movie Experience
Top Terms: 'movie,' 'film,' 'see,' 'one,' 'bad,' 'time,' 'even,' 'play,' 'say,' 'work.'
Description: This topic contains words that suggest dissatisfaction with films, such as 'bad,' 'see,' and 'time.' The frequent occurrence of 'movie' and 'film' suggests that these terms relate to negative opinions or experiences associated with watching movies.

Topic 1: Film Criticism
Top Terms: 'film,' 'movie,' 'bad,' 'watch,' 'even,' 'really,' 'didn't,' 'character,' 'story,' 'see.'
Description: The presence of terms like 'bad,' 'watch,' 'didn't,' and 'really' suggests this topic is about critical reviews or negative feedback regarding films and their characters or storylines.

Topic 2: Positive Film Review
Top Terms: 'film,' 'movie,' 'also,' 'good,' 'make,' 'one,' 'well,' 'script,' 'look,' 'great.'
Description: This topic focuses on positive comments about films, with words like 'good,' 'make,' and 'great' indicating approval. The mention of 'script' and 'look' suggests that people are praising the quality of the film's production or narrative.

Topic 3: Emotional Film Reactions
Top Terms: 'film,' 'love,' 'movie,' 'one,' 'bad,' 'good,' 'really,' 'see,' 'watch,' 'well.'
Description: Words like 'love,' 'bad,' 'good,' and 'really' indicate a passionate or emotional response to movies, suggesting a mix of positive and negative emotions.

Topic 4: Film Thoughts
Top Terms: 'film,' 'movie,' 'think,' 'well,' 'like,' 'good,' 'character,' 'give,' 'don't,' 'excellent.'
Description: This topic seems to describe thoughtful and reflective opinions, with terms like 'think,' 'like,' and 'don't,' suggesting a more constructive approach to film criticism.

Topic 5: Mixed Film Criticism
Top Terms: 'movie,' 'bad,' 'film,' 'see,' 'like,' 'make,' 'one,' 'really,' 'could,' 'well.'
Description: This topic conveys a combination of dissatisfaction ('bad,' 'really') and more neutral or mixed thoughts ('could,' 'well'). It is a critique of films, but with some positive aspects or qualifications.

Topic 6: Film Scenes and Acting
Top Terms: 'one,' 'movie,' 'scene,' 'character,' 'act,' 'film,' 'ever,' 'time,' 'also,' 'minute.'
Description: Words like 'scene,' 'character,' and 'act' suggest the topic revolves around the quality of film scenes, acting performances, and the overall cinematic experience.

Topic 7: Film Acting and Cast Review
Top Terms: 'movie,' 'act,' 'bad,' 'film,' 'cast,' 'make,' 'well,' 'like,' 'scene,' 'watch.'
Description: This topic seems to focus on the acting performances, with terms like 'act,' 'cast,' and 'scene' dominating. The combination of 'bad' suggests some critique, but the overall focus is on the performers and their contributions.

Topic 8: Character and Story Critique
Top Terms: 'movie,' 'film,' 'character,' 'like,' 'see,' 'bad,' 'good,' 'make,' 'would,' 'watch.'
Description: The presence of 'character' and 'story-related terms like 'good' and 'bad' suggest this topic is focused on opinions about the characters and overall narrative of the films.

Topic 9: Overall Movie Enjoyment
Top Terms: 'movie,' 'film,' 'one,' 'make,' 'time,' 'good,' 'great,' 'bad,' 'end,' 'still.'
Description: The frequent appearance of terms like 'good,' 'great,' and 'bad,' along with 'end,' suggests that this topic relates to the overall enjoyment of films, including their conclusion or final impression.

5.1.1 Quality Assessment of Detected Topics

The topics detected by LDA generally make sense, as they group together related words that reflect common themes in movie reviews or discussions. However, there are some overlaps, such as the presence of "bad," "good," "movie," and "film" in multiple topics, which might indicate that these terms have different weights across topics depending on context. The topics also seem to be biased toward film criticism, with a focus on positive and negative opinions, acting, and story elements as it is about a movie review. Overall, the topics are well-formed and provide meaningful insights into how people rate movies. OpenAI, 2025 Developers, 2025

5.2 Non-negative Matrix Factorization (NMF) (EXTRA)

The analysis of the 10 topics is done below.

Topic 1: Positive Movie Experience
Top Terms: 'movie,' 'stupid,' 'make,' 'love,' 'rate,' 'awesome,' 'way,' 'get,' 'great,' 'lot'
Description: This topic is related to positive comments about a movie, describing it with terms like 'awesome,' 'great,' and 'lot.' The word 'love' also suggests a favorable reaction to the film.

Topic 2: Praise for Film and Director
Top Terms: 'film,' 'great,' 'excellent,' 'make,' 'short,' 'saw,' 'director,' 'look,' 'truly,' 'christmas'
Description: This topic is about highly appreciating the film, its director, and its production quality, with terms like 'great,' 'excellent,' and 'director.' It also references Christmas, possibly indicating holiday-related films.

Topic 3: Negative Film Experience
Top Terms: 'bad,' 'act,' 'one,' 'even,' 'everything,' 'thought,' 'write,' 'series,' 'yes,' 'badwellits'
Description: This topic represents negative feedback or dissatisfaction with a film. Words like 'bad,' 'act,' and 'everything' indicate disappointment with the acting, story, or overall film quality.

Topic 4: Mixed Film Opinion
Top Terms: 'see,' 'one,' 'ever,' 'go,' 'definitely,' 'ive,' 'worth,' 'show,' 'anyone,' 'come'
Description: This topic is about a more neutral or mixed reaction to films, suggesting a balance of pros and cons, with words like 'worth,' 'definitely,' and 'show.'

Topic 5: Appreciation for Cast and Story
Top Terms: 'good,' 'cast,' 'great,' 'transfer,' 'quite,' 'end,' 'sad,' 'director,' 'actor,' 'love'
Description: This topic reflects a positive appreciation of the cast, story, and overall production. The terms 'good,' 'great,' and 'love' point to positive feedback about the movie's cast and emotional impact.

Topic 6: Criticism of Acting and Direction
Top Terms: 'like,' 'really,' 'look,' 'camera,' 'hate,' 'much,' 'funny,' 'anyone,' 'wonderful,' 'think'
Description: This topic covers mixed or critical opinions about the direction or acting in the film, with words like 'hate,' 'really,' and 'look,' indicating dissatisfaction or strong personal opinions.

Topic 7: Recommendations and Social Influence
Top Terms: 'recommend,' 'nothing,' 'highly,' 'friend,' 'saw,' 'confidence,' 'anyone,' 'definitely,' 'sorry,' 'cant'

Description: This topic is about sharing recommendations with others, possibly suggesting films to friends or family. The words like 'recommend,' 'highly,' and 'definitely' emphasize a favorable suggestion.

Topic 8: Predictable and Average Film
Top Terms: 'watch,' 'go,' 'keep,' 'easy,' 'get,' 'predictable,' 'omit,' 'thing,' 'even,' 'joy'
Description: This topic represents opinions on films that are seen as predictable or average, with terms like 'predictable,' 'easy,' and 'thing,' indicating a lack of excitement or originality.

Topic 9: Film Plot and Character Analysis
Top Terms: 'well,' 'plot,' 'character,' 'act,' 'script,' 'give,' 'played,' 'real,' 'write,' 'cast'
Description: This topic focuses on the plot, character development, and the quality of acting and script in a film, with terms like 'plot,' 'character,' 'act,' and 'script.'

Topic 10: Disappointment and Waste of Time
Top Terms: 'awful,' 'time,' 'dont,' 'waste,' 'even,' 'line,' 'story,' 'one,' 'worth,' 'look'
Description: This topic highlights strong dissatisfaction with films, with words like 'awful,' 'waste,' and 'dont,' emphasizing a negative experience where the film is considered a waste of time.

5.2.1 Quality Assessment of Detected Topics

The quality of the topics generated by NMF provides clear and interpretable themes, with each topic capturing distinct aspects of movie reviews, such as positive experiences, negative critiques, acting, direction, and emotional responses. The topics are well-separated with minimal overlap, making them easy to interpret. However, some topics, like Topic 4 (Mixed or Indifferent Film Opinion), could be more specific, as they contain a variety of terms that don't form a coherent narrative. Overall, the topics appear to cover a broad range of movie-related discussions, making NMF useful for understanding different aspects of film sentiment. OpenAI, 2025

5.3 Comparison Between Two methods

Compared to LDA (Latent Dirichlet Allocation), NMF tends to produce topics that are more focused and interpretable, resulted in more coherent terms. LDA, on the other hand, is more flexible but sometimes less straightforward. LDA generates topics with broader themes and less specificity, while NMF' can provide sharper, well-defined topics. It can also be seen in the Figure5.2 that the boundaries are more distant than each other in NMF rather than LDA shown in Figure5.1. OpenAI, 2025

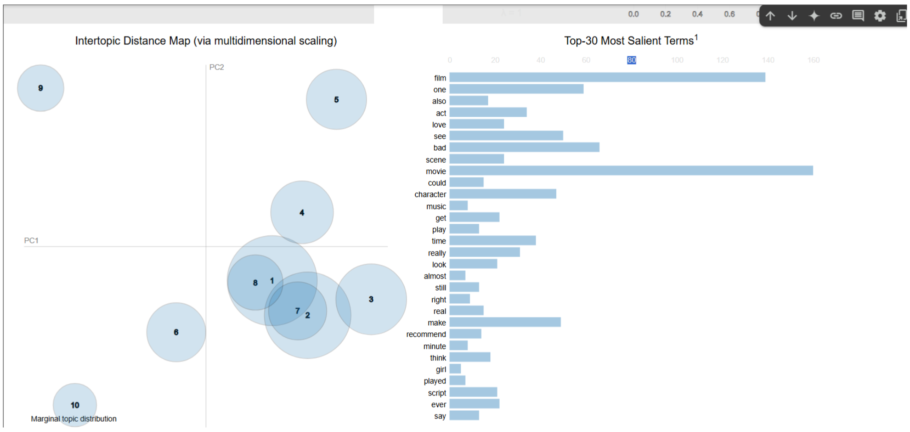


Figure 5.1: LDA Topic Detection

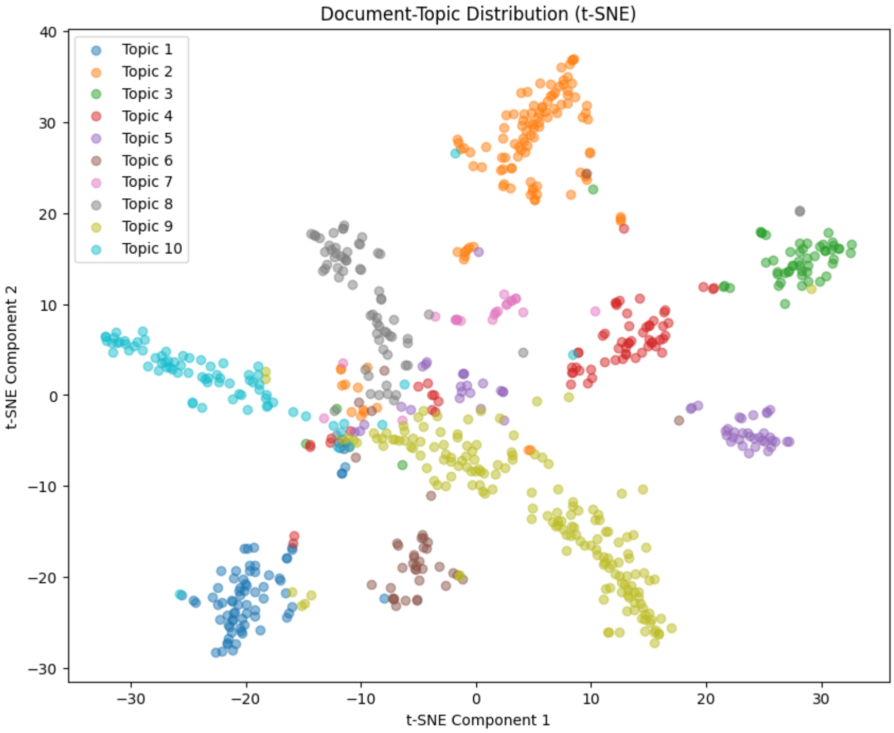


Figure 5.2: NMF Topic Detection

However, both methods can be effective for topic modeling, with NMF potentially offering more intuitive results.

References

Developers, S.-l. (2025). Scikit-learn: Machine learning in python [Accessed: 2025-01-22]. https://scikit-learn.org/

Mohasseb, D. A. (2025). Lecture slides for intelligent data and text analytics [Lecture slides for Intelligent Data and Text Analytics, University of Portsmouth].

OpenAI. (2025). Chatgpt: Language model [Accessed: 2025-01-22]. https://chat.openai.com/

[NOTE: ChatGPT was used to faciliate better understanding, and correcting sentence and grammatical errors.]