Intelligent Data and Text Analytics Coursework 2

Student Number: 2089114
Course: MSc Data Analytics
Year: 2024

Word count: 2,993

Table of Contents

1. Data Preprocessing

Data preprocessing is essential in natural language processing tasks, to clean and standardise the text. This makes machine learning models more suitable for extracting meaningful insights and improving overall model performance. Pre-processed data will be applied to Bag-of-Words classification and topic detection in this report, BERT-based model for classification utilises a built-in preprocessing model, so manual preprocessing is not necessary.

Order of preprocessing tasks can be seen in Figure 1.

Figure 1.
Sequence of Text Preprocessing Steps

1.1 *Removing punctuation*

Punctuation marks such as full stops, commas, question marks, brackets or any other special character are removed from the text. This cleans the data to focus on the actual text without any distracting or unnecessary symbols, see Table 1.1.

Table 1.1
Removing punctuation

Example	Original Text	Removing Punctuation
1	This totally UNfunny movie is so over the top and pathetic and unrealistic that throughout the whole 90 minutes of utter torture I probably looked at my watch about 70000 times!	This totally UNfunny movie is so over the top and pathetic and unrealistic that throughout the whole 90 minutes of utter torture I probably looked at my watch about 70000 times
2	Only like 3 or 4 buildings used, a couple of locations MAYBE, & poor hummh!	Only like 3 or 4 buildings used a couple of locations MAYBE poor hummh
3	As a European, the movie is a nice throwback to my time as a student in the 1980's and the experiences I had living abroad and interacting with other nationalities, although the circumstances were slightly different.	As a European the movie is a nice throwback to my time as a student in the 1980s and the experiences I had living abroad and interacting with other nationalities although the circumstances were slightly different
4	20th Century Fox's ROAD HOUSE 1948) is not only quite a silly noir but is an implausible unmitigated bore of a movie.	20th Century Foxs ROAD HOUSE 1948 is not only quite a silly noir but is an implausible unmitigated bore of a movie
5	The original Body and Soul (1947) is a masterpiece.	The original Body and Soul 1947 is a masterpiece
6	But "Tiny Toons" kept the 90's vibe and delivered one of the most popular, funny, and underrated cartoons ever created.	But Tiny Toons kept the 90s vibe and delivered one of the most popular funny and underrated cartoons ever created
7	I saw it as a child on TV back in 1973, when it was "The Stranger" and I loved it.	I saw it as a child on TV back in 1973 when it was The Stranger and I loved it
8	Still, it was the SETS that got a big "10" on my "oy-vey" scale.	Still it was the SETS that got a big 10 on my oyvey scale

1.2 *Removing numbers*
Any numerical value is removed from the text, to ensure the model focuses on textual content, avoiding noise introduced by numbers, although removing numbers may sometimes alter the meaning and interpretation of a review, such as example 8, removing the digits makes the text unclear.

Table 1.2
Text without Numbers

Example	Text after removing punctuation	Text without numbers
1	This totally UNfunny movie is so over the top and pathetic and unrealistic that throughout the whole **90** minutes of utter torture I probably looked at my watch about **70000** times	This totally Unfunny movie is so over the top and pathetic and unrealistic that throughout the whole minutes of utter torture I probably looked at my watch about times
2	Only like **3** or **4** buildings used a couple of locations MAYBE poor hummh	Only like or buildings used a couple of locations MAYBE poor hummh
3	As a European the movie is a nice throwback to my time as a student in the **1980**s and the experiences I had living abroad and interacting with other nationalities although the circumstances were slightly different	As a European the movie is a nice throwback to my time as a student in the s and the experiences I had living abroad and interacting with other nationalities although the circumstances were slightly different
4	**20**th Century Foxs ROAD HOUSE **1948** is not only quite a silly noir but is an implausible unmitigated bore of a movie	Th Century Foxs ROAD HOUSE is not only quite a silly noir but is an implausible unmitigated bore of a movie
5	The original Body and Soul **1947** is a masterpiece	The original Body and Soul is a masterpiece
6	But Tiny Toons kept the **90**s vibe and delivered one of the most popular funny and underrated cartoons ever created	But Tiny Toons kept the s vibe and delivered one of the most popular funny and underrated cartoons ever created
7	I saw it as a child on TV back in **1973** when it was The Stranger and I loved it	I saw it as a child on TV back in when it was The Stranger and I loved it
8	Still it was the SETS that got a big **10** on my oyvey scale	Still it was the SETS that got a big on my oyvey scale

1.3 *Changing all text to lowercase*
All text has been converted to lowercase, to ensure uniformity in the text, treating all words the same, see Table 1.3.

Table 1.3
Text all lowercase

Example	Text after removing numbers	Lowercase
1	**T**his totally **U**nfunny movie is so over the top and pathetic and unrealistic that throughout the whole minutes of utter torture **I** probably looked at my watch about times	this totally unfunny movie is so over the top and pathetic and unrealistic that throughout the whole minutes of utter torture i probably looked at my watch about times
2	**O**nly like or buildings used a couple of locations **MAYBE** poor hummh	only like or buildings used a couple of locations maybe poor hummh
3	**A**s a **E**uropean the movie is a nice throwback to my time as a student in	as a european the movie is a nice throwback to my time as a student in the

	the s and the experiences **I** had living abroad and interacting with other nationalities although the circumstances were slightly different	s and the experiences i had living abroad and interacting with other nationalities although the circumstances were slightly different
4	th **C**entury **F**oxs **ROAD HOUSE** is not only quite a silly noir but is an implausible unmitigated bore of a movie	th century foxs road house is not only quite a silly noir but is an implausible unmitigated bore of a movie
5	The original **B**ody and **S**oul is a masterpiece	the original body and soul is a masterpiece
6	But **T**iny **T**oons kept the s vibe and delivered one of the most popular funny and underrated cartoons ever created	but tiny toons kept the s vibe and delivered one of the most popular funny and underrated cartoons ever created
7	**I** saw it as a child on **TV** back in when it was **T**he **S**tranger and **I** loved it	i saw it as a child on tv back in when it was the stranger and i loved it
8	**S**till it was the **SETS** that got a big on my oyvey scale	still it was the sets that got a big on my oyvey scale

1.4 *Removing stop words*

Stop words refer to commonly used words (e.g. 'and,' 'the,' 'is'). Such words do not contribute meaningfully to text and therefore removed. Doing so reduces dimensionality and improves efficiency of machine learning models.

Table 1.4

Text no stop words

Example	Text after transformed to lowercase	Text No Stop Words
1	**this** totally unfunny movie **is so over the** top **and** pathetic **and** unrealistic **that throughout the** whole minutes **of** utter torture **i** probably looked **at my** watch **about** times	totally unfunny movie top pathetic unrealistic throughout whole minutes utter torture probably looked watch times
2	**only** like **or** buildings used **a** couple **of** locations maybe poor hummh	like buildings used couple locations maybe poor hummh
3	**as a** european **the** movie **is a** nice throwback **to my** time **as a** student **in the s and the** experiences **i had** living abroad **and** interacting **with** other nationalities although **the** circumstances **were** slightly different	european movie nice throwback time student experiences living abroad interacting nationalities although circumstances slightly different
4	th century foxs road house **is not only** quite **a** silly noir **but is an** implausible unmitigated bore **of a** movie	th century foxs road house quite silly noir implausible unmitigated bore movie
5	**the** original body **and** soul **is a** masterpiece	original body soul masterpiece
6	**but** tiny toons kept **the s** vibe **and** delivered one **of the most** popular funny **and** underrated cartoons ever created	tiny toons kept vibe delivered one popular funny underrated cartoons ever created

7	**i** saw **it as a** child **on** tv back **in when it was the** stranger **and i** loved **it**	saw child tv back stranger loved
8	still **it was the** sets **that** got **a** big **on my** oyvey scale	still sets got big oyvey scale

1.5. *Tokenisation*
This technique splits text into individual words (tokens). This step is essential for further processing, as it enhances the quality of machine learning models, including classification and topic detection.

Table 1.5
Text tokenised

Example	Text after removing stop words	Text Tokenised
1	totally unfunny movie top pathetic unrealistic throughout whole minutes utter torture probably looked watch times	totally, unfunny, movie, top, pathetic, unrealistic, throughout, whole, minutes, utter, torture, probably, looked, watch, times
2	like buildings used couple locations maybe poor hummh	like , buildings ,used, couple ,locations , maybe, poor, hummh
3	european movie nice throwback time student experiences living abroad interacting nationalities although circumstances slightly different	european, movie, nice, throwback, time, student, experiences, living, abroad, interacting, nationalities, although, circumstances, slightly, different
4	th century foxs road house quite silly noir implausible unmitigated bore movie	th, century, foxs, road, house, quite, silly, noir, implausible, unmitigated, bore, movie
5	original body soul masterpiece	original, body, soul, masterpiece
6	tiny toons kept vibe delivered one popular funny underrated cartoons ever created	tiny, toons, kept, vibe, delivered, one, popular, funny, underrated, cartoons, ever, created
7	saw child tv back stranger loved	saw, child, tv, back, stranger, loved
8	still sets got big oyvey scale	still, sets, got, big,oyvey,scale

1.5 *Lemmatising*
Words are reduced to their base or dictionary form, lemmatisation accounts for the context and parts of speech of words, this had very little difference on the examples used since majority of words/tokens were already in their base form.

Table 1.5
Text with lemmatising

Example	Text after tokenisation	Text with Lemmatising
1	**totally**, unfunny, movie, top, pathetic, unrealistic, throughout, whole, minutes, utter, torture, probably, looked, watch, times	total, unfunny, movie, top, pathetic, unrealistic, throughout, whole, minutes, utter, torture, probably, look, watch, times
2	like, buildings , **used**, couple ,locations , maybe, poor, hummh	like, buildings, use, couple, locations, maybe, poor, hummh

3	european, movie, nice, throwback, time, student, experiences, **living**, abroad, **interacting**, **nationalities**, although, circumstances, **slightly**, different	european, movie, nice, throwback, time, student, experiences, liv, abroad, interact, nationaliti, although, circumstances, slight, different
4	th, century, fox**s**, road, house, quite, silly, noir, implausible, **unmitigated**, bore, movie	th, century, fox, road, house, quite, silly, noir, implausible, unmitigate, bore, movie
5	original, body, soul, masterpiece	original, body, soul, masterpiece
6	tiny, toons, kept, vibe, **delivered**, one, popular, funny, underrated, cartoons, ever, **create**	tiny, toons, kept, vibe, deliver, one, popular, funny, underrated, cartoons, ever, created
7	saw, child, tv, back, stranger, **loved**	saw, child, tv, back, stranger, love
8	still, sets, got, big, oyvey, scale	still, sets, got, big, oyvey, scale

1.6. *Stemming*
This is similar to lemmatising, but words are reduced to their base or dictionary form by removing suffixes, no matter the grammatical context. It is less precise than stemming but a much faster alternative. Like with lemmatising, there was very little difference in the text here. Lemmatising already reduced words to its base form, stemming further simplifies text that lemmatising does not, primarily by removing tense and plural suffixes:

Table 1.6
Text with stemming

Example	Text after lemmatisation	Text with Stemming
1	total, unfunny, movie, top, pathetic, unrealistic, throughout, whole, **minutes**, utter, torture, **probably**, look, watch, **times**	total, unfunny, movie, top, pathetic, unrealistic, throughout, whole, minute, utter, torture, probabl, look, watch, time
2	like, **buildings**, use, couple, **locations**, maybe, poor, hummh	like, building, use, couple, location, maybe, poor, hummh
3	european, movie, nice, throwback, time, student, **experiences**, live, abroad, interact, nationality, although, **circumstances**, slight, different	european, movie, nice, throwback, time, student, experience, live, abroad, interact, nationality, although, circumstance, slight, different
4	th, century, fox, road, house, quite, silly, noir, implausible, unmitigate, bore, movie	th, century, fox, road, house, quite, silly, noir, implausible, unmitigate, bore, movie
5	original, body, soul, masterpiece	original, body, soul, masterpiece
6	tiny, toons, kept, vibe, deliver, one, popular, funny, underrat**ed**, **cartoon**, ever, create	tiny, toons, kept, vibe, deliver, one, popular, funny, underrat, cartoons, ever, create
7	saw, child, tv, back, stranger, love	saw, child, tv, back, stranger, love
8	still, **sets**, got, big, oyvey, scale	still, set, got, big, oyvey, scale

Now that the data has been pre-processed, it is ready to be used for classification and topic detection.

Task 2: BoW Classification

2.1 Data Preparation

Data Transformation

- The IMDb reviews dataset was transformed using the TF-IDF Vectoriser, converting text into numerical representation by calculating the *term frequency-inverse document frequency.* This creates a Bag-of-Words representation for classification.

Train-Test Splitting and Balancing

- The reviews data was split into training and testing sets required for classification models.
- Slicing technique was used to balance the train and test sets, the training set included the first 500 samples, while the test set used the remaining 248 samples:

```
#train dataset by splitting the data
train_reviews = reviewsP.Review[:500]
train_sentiments = reviewsP.Sentiment[:500]

#test dataset
test_reviews = reviewsP.Review[248:]
test_sentiments = reviewsP.Sentiment[248:]

print(train_reviews.shape,train_sentiments.shape)
print(test_reviews.shape,test_sentiments.shape)
```

```
(500,) (500,)
(500,) (500,)
```

2.2 Classification with Naïve Bayes:

Table 2.2a.
Performance metrics for Naïve Bayes

	Precision	Recall	F1-score	Support	Overall Accuracy	Overall AUC
Positive	0.76	0.97	0.85	225	0.852	0.863
Negative	0.97	0.76	0.85	275		

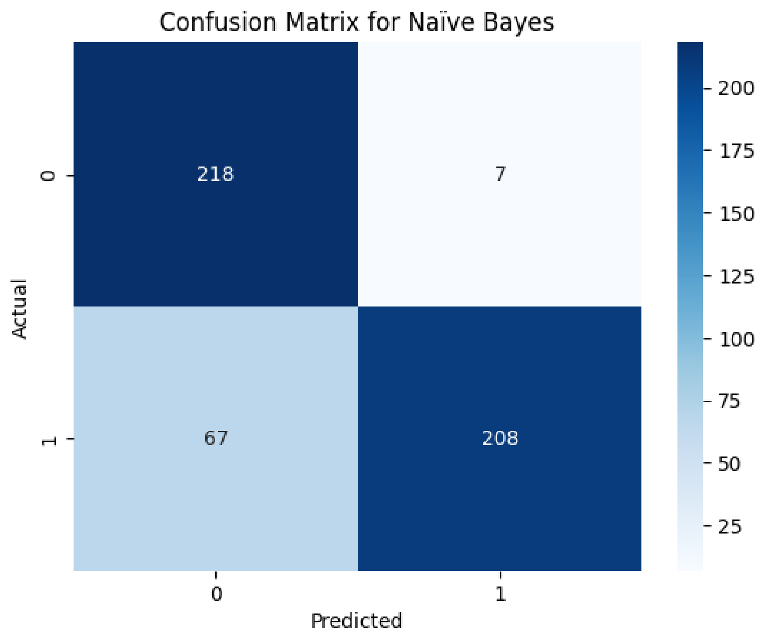
Positive Reviews: Precision score demonstrates that out of all predicted positives, 76% were true positive predictions. A recall of 97% means all actual positive reviews were correctly identified.

Negative Reviews: Precision indicates NB accurately predicted 97% negative instances, outperforming precision of positive reviews in accuracy of identifying correct sentiments. Recall of 0.76 shows NB captured 76% of actual negative instances, recall was higher for positive reviews, suggesting NB exhibited greater sensitivity to positive sentiments.

Overall: Across all instances, 85.20% of predictions were accurately predicted, while an AUC of 86.30% indicates great ability of the model to distinguish between both classes. Overall, NB performed strongly.

Table 2.2b.
Error metrics for NB

Number of instances	True N	False N	False P	True P	Total Correct Predictions	% of correct predictions	Total incorrect Predictions	% of incorrect predictions
500	218	67	7	208	426	85.20%	74	14.80%



This model performed well, correctly predicting 85.20% instances: 218 positive and 208 negatives reviews. Although, there is still room for improvement, given 14.80% of instances were not correctly predicted.

Table 2.2c.

Evaluation of NB

Strengths	Weaknesses
Efficiency: this model is computationally efficient and provides quick predictions, great for larger datasets	**Feature independence:** This assumes features, or text characteristics such as word tokens, are independent, which might not always be true in real-world datasets.
Simplicity: the simplicity of the Naïve Bayes classifier makes it easy to implement and a great baseline model that requires minimal training time	**Poorer performance in complex relationships:** where classes show greater complexity, Naïve Bayes does not reveal interactions between features as effectively
Performance on text data: it often performs very well on text classification tasks, where Naïve Bayes' assumption of feature independence can be reasonable, but this is also a weakness.	

2.3 Classification with K-Nearest Neighbour

Table 2.3a

Performance metrics for KNN

	Precision	Recall	F1-score	Support	Overall Accuracy	Overall AUC
Positive	0.77	0.71	0.74	225	0.774	0.768
Negative	0.78	0.83	0.80	275		

Positive Reviews: KNN correctly identified 77% positive sentiments out of all positive predictors. A recall of 0.71 means the model captured 71% of all actual positive instances. A F1-score of 0.74 shows the model moderately predicted positive reviews while minimising false negatives and positives.

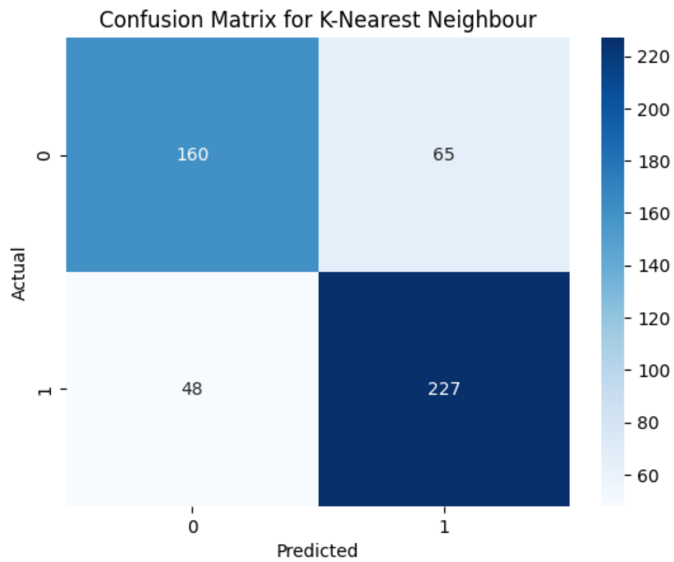
Negative Reviews: 78% of reviews were correctly predicted as negatives. KNN captured 83% of actual negative reviews, showing greater sensitivity to the negative sentiments than positive. F1-score of 0.80 shows very good balance between precision and recall, slightly outperforming the positive class.

Overall performance: Overall model accuracy is 77.4%, this is relatively a good score but there is still room for improvement, particularly in recall scores for the positive sentiment. An AUC of 0.768 demonstrates moderate performance in distinguishing between positive and negative reviews.

Table 2.3b

Error metrics for KNN

Number of instances	True N	False N	False P	True P	Total Correct Predictions	% of correct predictions	Total incorrect Predictions	% of incorrect predictions
500	160	48	65	227	387	77.40%	113	22.60%



KNN performed moderately, correctly predicting 77.40% sentiments, although it identified a reasonable proportion of incorrect predictions (22.60%): 48 false negatives, 65 false positives; suggesting there is a lot of room for improvement.

Table 2.3c

Evaluation of KNN

Strengths	Weaknesses
Simplicity: KNN is simple and easy to understand, deeming it appropriate for quick implementations in classification tasks.	**Requires a lot of memory:** KNN stores entire training dataset, so it can be memory-intensive, particularly with larger datasets.
Performs well in complexity: where decision boundaries are complex, KNN can adapt to it, where there is no linearity between decision boundaries, KNN can still make effective predictions.	**Sensitive to irrelevant features:** distance between instances can be distorted by features that are not of useful information

2.4 Classification with Support Vector Machine

Table 2.4a

Performance metrics for SVM

	Precision	Recall	F1-score	Support	Overall Accuracy	Overall AUC
Positive	0.81	0.92	0.86	225	0.866	0.871
Negative	0.93	0.82	0.87	275		

Positive reviews: 81% of instances were correctly predicted as positive out of all positive predictors. 92% were identified as acutal positive reviews. A F1-Score 0f 0.86 demonstrates the model balances both false positives and false negatives effectively.

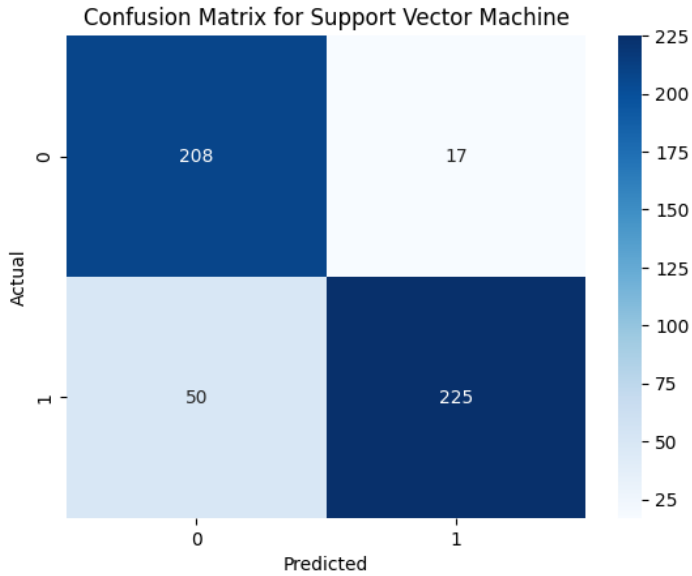
Negative reviews: 93% of sentiments were correctly predicted as negative, a recall of 0.82 indicates 82% were actual negative reviews, giving room for improvement in correctly classifying negative reviews. A similar F1-score to positive instances indicates this model performed well in predicting negative reviews while minimising false negatives and positives.

Overall: SVM made 86.6% accurate sentiment predictions across both classes, a strong performance. An AUC of 0.871 shows the model performed well in distinguishing between both classes.

Table 2.4b

Error metrics for SVM

Number of instances	True N	False N	False P	True P	Total Correct Predictions	% of correct predictions	Total incorrect Predictions	% of incorrect predictions
500	208	50	17	225	433	86.60%	67	13.40%



The SVM model performed well in correctly identifying 433 instances (86.60%) while 67 were incorrectly predicted (13.40%). While there is still room for improvement, this model still performed strongly.

Table 2.4c

Evaluation of SVM

Strengths	Weaknesses

Effective in high-dimensional spaces: SVM thrives in situations with many features, such as text classification.	**Training time**: computationally expensive and slow, the training time drastically increases as the dataset size increases.
Prone to overfitting: robust to overfitting, particularly in high-dimensional spaces.	**Performance limitations:** SVM can struggle with imbalanced datasets - might show bias towards the majority class.
Use of Kernels: this allows SVM to handle non-linear data as well as linear.	

2.5 Classification with Decision Tree

Table 2.5a
Performance metrics for DT

	Precision	Recall	F1-score	Support	Overall Accuracy	Overall AUC
Positive	0.81	0.87	0.84	225	0.850	0.852
Negative	0.88	0.84	0.86	275		

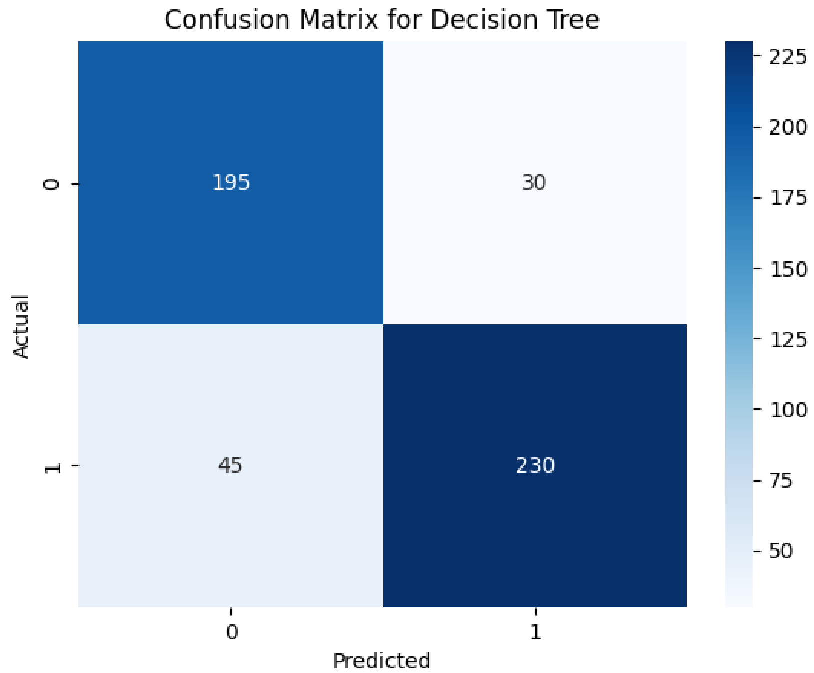
Positive Reviews: Of all positive predictors, 81% were accurately identified as a positive review. Recall demonstrates 87% of all actual positive reviews were correctly classified. A F1-score of 0.84 indicates good balance between precision and recall.

Negative Reviews: The DT model performed better in accurately predicting negative instances than positive instances, given a higher precision value (0.88). A recall of 0.84 shows 84% were actual negative instance. F1-score is only marginally higher for negative reviews than positive, also indicating great balance between precision and recall.

Overall: An overall accuracy of 85% is relatively high, while an AUC of 85.20% suggests the model performed well in distinguishing between both positive and negative classes. The DT model performed well overall.

Table 2.5b
Error metrics for DT

Number of instances	True N	False N	False P	True P	Total Correct Predictions	% of correct predictions	Total incorrect Predictions	% of incorrect predictions
500	195	45	30	230	425	85.00%	75	15.00%



DT model correctly predicted 425 instances (85%), while incorrectly predicting 75 (15%), 45 instances were positive but predicted as negative, while 30 instances were negative but classified as positive. This model performed very well, but improvements can be implemented to reduce the error.

Table 2.5c

Evaluation of DT

Strengths	Weaknesses
Simplicity and interpretability: Decision Trees are easy to implement, and understand	**Overfitting:** prone to overfitting, especially when the depth of the tree increases.
Handles non-linear relationships: Decision Trees can model more complex relationships.	**Bias towards majority:** where data is imbalanced, decision trees show bias towards the majority class.
Training time: not as computationally expensive as other models, even with larger datasets.	

2.6 Classification with Logistic Regression

Table 2.6a

Performance metrics for LR

	Precision	Recall	F1-score	Support	Overall Accuracy	Overall AUC
Positive	0.77	0.96	0.85	225	0.854	0.863
Negative	0.95	0.77	0.85	275		

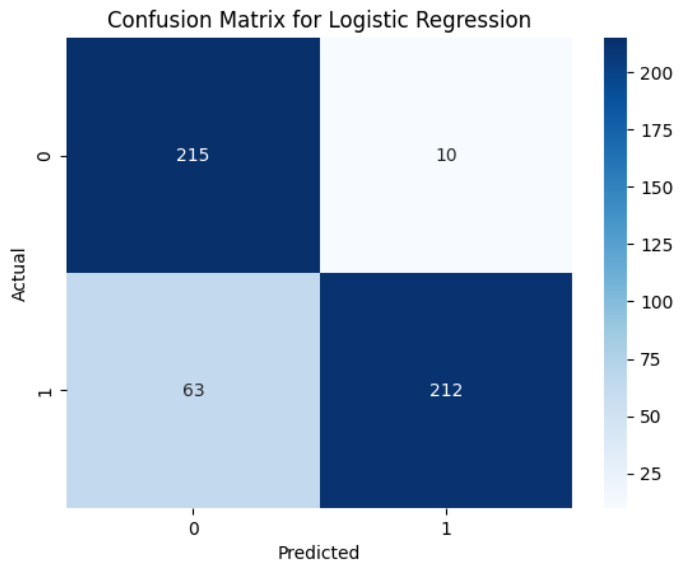
Positive reviews: 77% of positive classes were correctly predicted out of all positive predictors, a recall of 0.96 indicates LR captured 96% of actual positive instances. An F1-score of 0.85 is high and demonstrates good balance between precision and recall.

Negative reviews: Of all negative predictors, 95% of negative reviews were correctly predicted, although a recall of 0.77 demonstrates only 77% of predictions were actual negative instances. An F1-score of 0.85 demonstrates balance between precision and recall.

Overall: An overall accuracy of 85.40% highlights LR's strong performance. An AUC of 86.30 shows great discriminatory power of the model in distinguishing between positive and negative classes. Although, adjustments can be made to refine the precision of positive reviews and the recall of negative reviews.

Table 2.6b
Error metrics for LR

Number of instances	True N	False N	False P	True P	Total Correct Predictions	% of correct predictions	Total incorrect Predictions	% of incorrect predictions
500	215	63	10	212	427	85.40%	73	14.60%



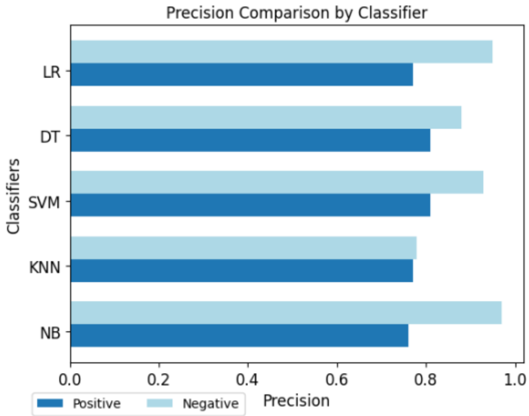
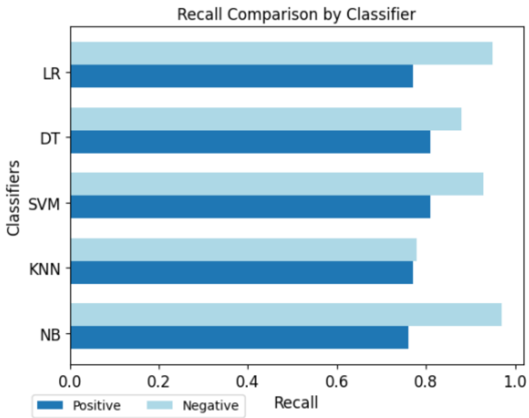
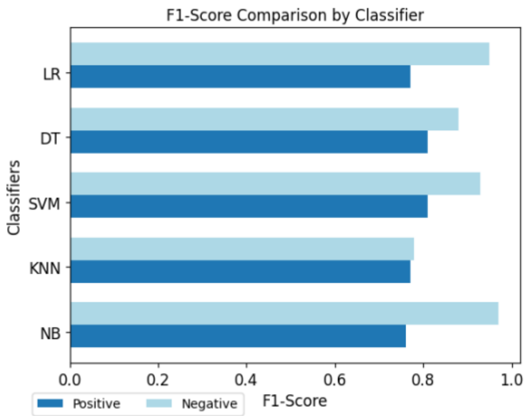
427 instances were correctly predicted (85.40%), a strong performance, although 14.60% of instances were misclassified, with 63 cases being false negatives, and 10 cases appeared to be false positives.

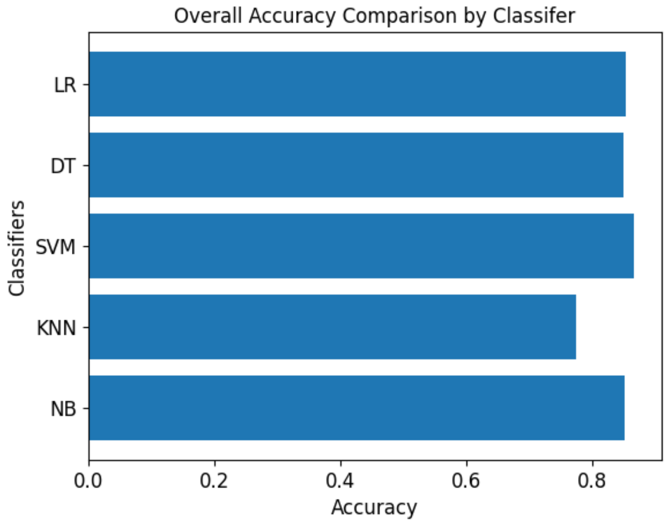
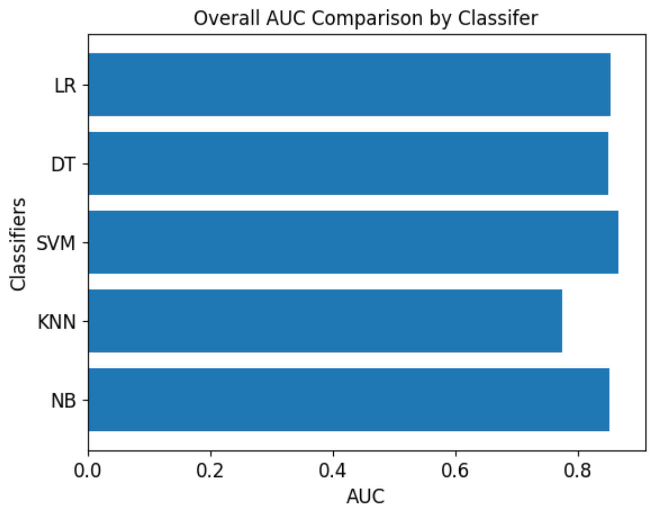
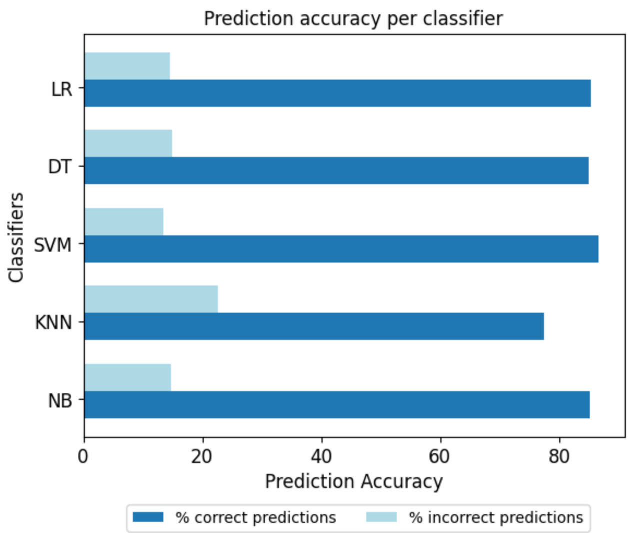
Table 2.6c
Evaluation of LR

Strengths	Weaknesses
Simplicity and interpretability: LR is easy to implement and understand. It provides insights into relationships between features and target variable.	**Limited application:** the assumption of a linear relationship between independent variables and log-odds of target features limits the application of LR to complex, non-linear relationships
Linear relationships: LR operates on assumptions of linearity, so in cases where there is linear relationship between features and targets, this model performs exceptionally well	**Bias towards majority:** where data is imbalanced, decision trees show bias towards the majority class.
Efficiency: computationally efficient, works well for small datasets.	**Performance limitations:** where there is high-dimensional data, LR may underperform

2.7 Comparison

Table 2.7
Table of comparison for all classifiers

Comparison Chart	Interpretation
	LR, SVM, NB performed the best at correctly predicting negative instances, DT performed well, but moderately compared to other models, KNN performed the worst. DT and SVM performed the best in precision for positive instances, followed by KNN, NB then LR, but these differences in precision scores are only marginally different.
	LR, SVM, NB performed the best at identifying actual negative instances with fewest false negatives, DT performed well, but moderately compared to other models, KNN performed the worst. DT and SVM performed the best in recall for positive instances, followed by LR, KNN then NB, but these differences in recall scores are only marginally different.
	LR and NB and SVM produced the greatest balance between precision and recall for negative classes. DT performed moderately, while KNN poorly balanced precision and recall. For positive instances, differences in F1-score were small, but DT and SVM balanced precision and recall of positive instances just as well as each other, marginally followed by LR, then KNN and NB.

	Overall accuracy of models is very similar. KNN performed the worst, LR, DT and NB all performed just as well as each other, but SVM takes the edge by a very slight difference. Nevertheless, LR, DT and NB performed just as strongly,
	KNN performed the worst at distinguishing between positive and negative classes. LR, DT and NB exhibited similar performances with only decimal differences between their AUC scores. SVM slightly outperforms these three classifiers, giving this model the greatest discriminatory power.
	Once again, KNN performed the worst, with the fewest correct predictions, and greatest incorrect predictions. LR, DT and NB performed very similarly, identifying a similar proportion of correct and incorrect predictions. SVM holds the greatest prediction accuracy, with the most correct and least incorrect predictions.

All classifiers performed well, KNN scores are still relatively good, but it was outperformed by all other models across all metrics, therefore it is the least suitable model for this dataset. The remaining models: LR, DT, NB and SVM performed exceptionally well, all achieving the greatest accuracy in correct predictions. However, across all metrics, LR, DT and NB performed very similarly with the smallest and most marginal differences between them. SVM consistently yielded the best results per metric, only slightly outperforming LR, DT and NB, deeming SVM the most suitable classification technique for this dataset, nevertheless, LR, DT and NB would still be appropriate in practice too.

Task 3: BERT Classification with Fine-Tuning

3.1 Fine-Tuning

To fine-tune the BERT-based model, these steps were conducted:

Optimising: the AdamW optimiser was generated with an initial learning rate of 3e-5, it also calculated warm-up steps (10% of total training steps), and total training steps consisted of 6 epochs, with 448 steps per epoch.

Data splitting: the dataset was split into training (X_train, y_train), test sets (X_test, y_test) and validation (X_val, y_val) to avoid biased evaluation.

```
from sklearn.model_selection import train_test_split

#split and take the test set
X, X_test, y, y_test = (train_test_split(reviews['Review'], reviews['Sentiment'],
                                          stratify=reviews['Sentiment'],
                                          test_size=0.2,
                                          train_size=0.8
                                          ))
```

```
#split the train set (X) into train and validation
X_train, X_val, y_train, y_val = (train_test_split(X, y,
                                                    stratify=y,
                                                    test_size=0.25,
                                                    train_size=0.75
                                                    ))
```

3.2 Classification with BERT

Table 3.2a

Performance metrics

	Precision	Recall	F1-Score	Support	Accuracy	AUC
Positive	0.64	0.68	0.66	73	0.660	0.661
Negative	0.68	0.64	0.66	77		

Positive reviews: The BERT-based model captured only 64% of correct predictions. A recall of 0.68 shows BERT identified 68% of actual positive instances. Precision and recall performance is moderate. An F1-score of 0.66 suggests moderate balance between precision and recall.

Negative reviews: Precision of negative instances slightly outperforms precision of positive instances by 0.04%. BERT identified 68% of correct negative predictions across all negative predictors. 64% are actual negative reviews, with an identical F1-score to positive instances. The BERT-based model performed moderately for classifying negative classes also.

Overall: Accuracy is also moderate, with 66% accurate predictions, and an AUC of 66.10% demonstrates a modest discriminatory power of this model in distinguishing between both classes.

Table 3.2b

Error metrics

Number of instances	True N	False N	False P	True P	Total Correct Predictions	% of correct predictions	Total incorrect Predictions	% of incorrect predictions
150	50	28	23	49	99	66.00%	51	34.00%

BERT correctly predicted 99 instances (66%), leaving a large proportion of predictions misclassified (34%), including 28 false negatives and 23 false positives. BERT model did not perform relatively well, requiring substantial improvements and refinements to reduce misclassifications.

Table 3.2c

Evaluation

Strengths	**Weakness**
Performance: BERT usually performs exceptionally across natural language processing tasks.	**High computational cost:** BERT-based models require a lot of computational resources and processing power, it is quite a slow algorithm due to this.
Minimal preprocessing: BERT includes built-in preprocessing models, making it less labour intensive to prepare data for BERT tasks	**Overfitting:** there is a risk of overfitting on small datasets.

3.3 Comparison to classifiers in section 2

Table 3.3a

Comparison against section 2 classifiers

	BERT with Fine-Tuning	Naïve Bayes	K-Nearest Neighbour	Support Vector Machine	Decision Tree	Logistic Regression
Precision P	0.64	0.76	0.77	0.81	0.81	0.77
Precision N	0.68	0.97	0.78	0.93	0.88	0.95
Recall P	0.68	0.97	0.71	0.92	0.87	0.96
Recall N	0.64	0.76	0.83	0.82	0.84	0.77
F1-Score P	0.66	0.85	0.74	0.86	0.84	0.85
F1-Score N	0.66	0.85	0.80	0.87	0.86	0.85
Overall accuracy	0.660	0.852	0.774	0.866	0.850	0.854
AUC	0.661	0.863	0.768	0.871	0.852	0.863

Number of instances	150	500	500	500	500	500
True P	50	218	160	208	195	215
False P	28	67	48	50	45	63
False N	23	7	65	17	30	10
True N	49	208	227	225	230	212
Total Correct Predictions	103	426	387	433	425	427
Percentage of total correct predictions	66.00%	85.20%	77.40%	86.60%	85.00%	85.40%
Total incorrect predictions	51	74	113	67	75	73
Percentage of total incorrect predictions	34.00%	14.80%	22.60%	13.40%	15.00%	14.60%

Across all metrics, BERT with fine-tuning has been outperformed by every model, even KNN, which had the poorest performance across previous models. BERT displays the worst precision, therefore identified the least number of correct predictions across both negative and positive classes; the worst recall, meaning it struggled the most in identifying actual negative and positive instances; and a low F1-score, signifying the worst harmony between precision and recall relative to other classifiers. Overall accuracy scores across other metrics quite distinctly outperform BERT, emphasising this model's poor performance in correctly predicting sentiments. A modest AUC value of 0.661 questions the discriminatory power of this model, relative to all other higher scoring models in comparison. This questions BERT's ability in distinguishing between both positive and negative classes. The confusion matrix emphasises this model's lack of accuracy, with the lowest percentage of total correct predictions and largest proportion of incorrect predictions.

Overall, BERT is the least suited for this dataset, perhaps due to a small number of instances, where BERT models tend to perform better on larger scale natural language processing tasks. However, BERT is a powerful classification model, so it was attempted to improve its performance using ensemble learning. Here, previous models from section 2 were combined with BERT to attempt to improve its overall performance. It is expected that given SVM performed the best in section 2, this will lead to the biggest improvement in BERT with fine-tuning performance, see Table 3.3b.

Table 3.3b

Overall performance and error metric comparison across ensemble techniques

	BERT with Fine-Tuning	**BERT + NB**	**BERT + KNN**	**BERT + SVM**	**BERT + DT**	**BERT + LR**
Precision P	0.64	**0.59**	**0.59**	**0.71**	**0.63**	**0.76**
Precision N	0.68	**0.73**	**0.62**	**0.72**	**0.72**	**0.78**
Recall P	0.68	**0.82**	**0.60**	**0.70**	**0.77**	**0.77**
Recall N	0.64	**0.45**	**0.61**	**0.73**	**0.57**	**0.77**
F1-Score P	0.66	**0.69**	**0.60**	**0.70**	**0.69**	**0.76**
F1-Score N	0.66	**0.56**	**0.61**	**0.72**	**0.64**	**0.77**
Overall accuracy	0.660	**0.633**	**0.607**	**0.713**	**0.600**	**0.767**

AUC	0.661	**0.638**	**0.607**	**0.713**	**0.601**	**0.767**
Number of instances	150	**150**	**150**	**150**	**150**	**150**
True P	50	**60**	**44**	**51**	**48**	**56**
False P	28	**42**	**30**	**21**	**35**	**18**
False N	23	**13**	**29**	**22**	**25**	**17**
True N	49	**35**	**47**	**56**	**42**	**59**
Total Correct Predictions	103	**95**	**91**	**107**	**90**	**427**
Percentage of total correct predictions	66.00%	**63.33%**	**60.67%**	**71.33%**	**60.00%**	**76.67%**
Total incorrect predictions	51	**55**	**59**	**43**	**60**	**35**
Percentage of total incorrect predictions	34.00%	**36.67%**	**39.33%**	**28.67%**	**40.00%**	**23.33%**

As predicted, when BERT combined with SVM, it yielded a great improvement achieving 71.33% accuracy, substantially better than its performance alone (66%). Although, this wasn't the greatest improvement, as the BERT-model paired with the Logistic Regressor resulted in the most notable improvement (76.67%). However, pairings with KNN, NB and DT produced mixed results, with accuracies ranging from 60-63% - not all ensemble combinations significantly enhanced BERT model results. Of interest, BERT paired with SVM and LR consistently outperformed all other models in both precision and recall, identifying the fewest false positives and negatives. Their F1-scores outweigh all models, but particularly BERT-model alone, conveying the greatest harmony between precision and recall.

In conclusion, BERT alone performed moderately, but these findings highlight the value of ensemble techniques. Using robust models such as SVM and LR, the BERT models performance enhanced its ability in delivering more reliable predictions.

4. Topic Detection

Topic detection identifies and extracts themes/topics from a large collation of unstructured text data, this will be performed using 3 algorithms:

1. BERT-Topic
2. BERT-Topic with K-Means (K-Means)
3. Non-Negative Matrix Foundation (NMF)

4.1 BERT-Topic:

Table 4.1a

Review and interpretation of topics and topic words

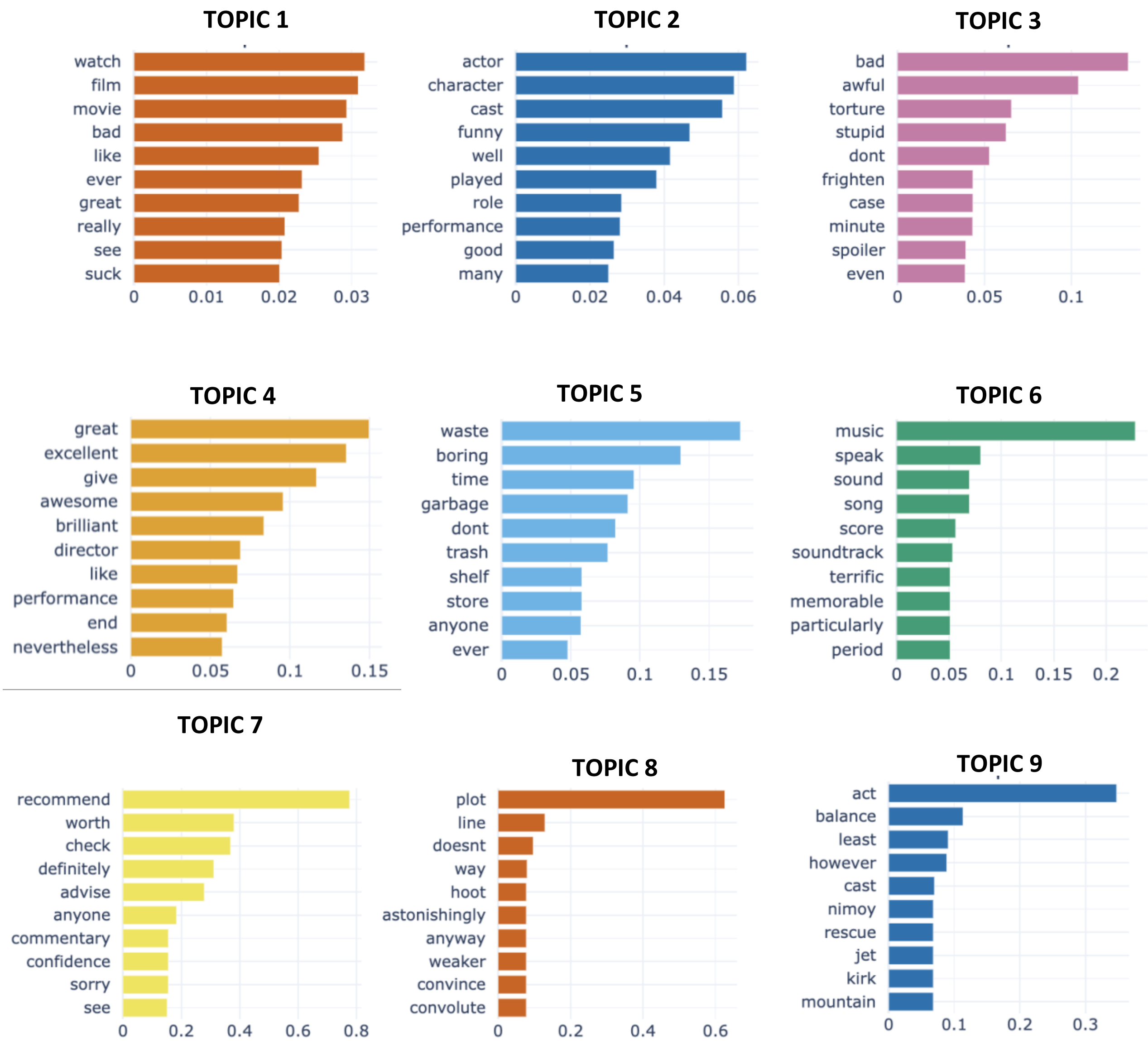
Topic	Top 10 Words	Top-Word interpretation	Interpretation
0	good, bad, character, script, scene, look, time, like, make, even	**Good:** this topic likely reflects general positive opinions within an IMDb review.	This topic implies mixed reviews, with the viewer mentioning good and bad aspects of the film/characters, script and specific scenes.
1	watch, film, movie, bad, like, ever, great, really, see, suck	**Watch:** topic 1 will may focus on the overall viewing experience.	Focus on general opinions of the film, some praise to the movie but also a strong dislike, using strong negative adjectives such as ‘suck.’
2	actor, character, cast, funny, well, played, role, performance, good, many	**Actor:** topic 2 will focus on a review of an actor	Emphasises the acting performances and character portrayal in a positive light.
3	bad, awful, torture, stupid, dont, frighten, case, minute, spoiler, even	**Bad:** topic 3 will strongly criticise a movie	A strong negative review of a film, viewer seems incredibly unsatisfied.
4	great, excellent, give, awesome, brilliant, director, like, performance, end, nevertheless	**Great:** topic 4 will strongly praise a movie	Exceptional reviews, the user highly praised the movie and director’s work.
5	waste, boring, time, garbage, don’t, trash, shelf, store, anyone, ever	**Waste:** topic 5 likely reflects a topic of a viewer expressing disappointment and dissatisfaction in the film.	Viewer considered the movie a waste of time and money.
6	music, speak, sound, song, score, soundtrack, terrific, memorable, particularly, period	**Music:** focus on the soundtrack and music used	Focuses on music and sound, capturing a positive response towards soundtracks and scores used in the film
7	recommend, worth, check, definitely, advise, anyone, commentary, confidence, sorry, see	**Recommend:** strong viewer endorsements and encouragement for others to watch	The viewer recommends this film, considers it worthwhile to watch, confident in the recommendation too.
8	plot, line, doesnt, way, hoot, astonishingly, anyway, weaker, convince, convolute	**Plot:** topic 8 likely discusses the storyline and structure of a film	Negative review that criticises the plot of the film as weak, convoluted and unconvincing.

9	act, balance, least, however, cast, nimoy, rescue, jet, kirk, mountain	**Act:** topic 9 centres around acting and character portrayal.	Focus of this topic is on specific features and characters of a film, but it is not clear if this is a positive or negative review.

Figure 3

Visual representation of top 10 words per topic with word weight

*** Note: No bar chart is presented for topic 0 as it is an outlier, but the top 10 words in heaviest weight distribution to lightest is: 'good,' 'bad,' 'character,' 'script,' 'scene,' 'look,' 'time,' 'like,' 'make,' and 'even.' ***



4.1b Sentiment Discussion using Figure 3:

Positive sentiments (Topic 2, 4 and 6, 7):
These charts feature words such as ‘excellent,’ ‘awesome,’ ‘terrific,’ ‘recommend,’ and ‘memorable,’ all of which are strongly weighted, indicating enjoyment and appreciation of films from the viewers, particularly in aspects such as actors’ performance and soundtracks.

Negative Sentiments (Topic 3 and 5):
These topic charts demonstrate largely weighted negative words such as ‘bad,’ ‘torture,’ and ‘waste.’ The strong weight for these negative terms indicates dissatisfaction from viewers.

Neutral Sentiments (Topic 0, 1 and 8): topics consisting of words such as ‘good,’ ‘bad,’ ‘watch’ and ‘plot’ are more evenly distributed in weight, revealing mixed and neutral opinions.

4.1c Quality Assessment

Topic coherence measures semantic similarity between the words in each topic. A higher score generally means the topics are more meaningful and interpretable if the words in the topic are more semantically similar. For the BERT-Topic model, the coherence score is **0.442**. This is a moderate score, providing both strengths and limitations of this model:

Strengths of BERT-Topic	Weakness of BERT-Topic
Moderate coherence: this score is not high but not low, its score is acceptable for interpretability, indicating that the topics have some meaning and cohesivity, but for optimal clarity refinement is needed.	**Moderate coherence:** this score is not high enough to determine all topics are entirely meaningful, some terms may be unrelated or loosely related, effecting the quality of the overall model.
Distinctiveness and Relevance: Some topics are coherent while focusing on certain elements of films that are usually discussed in IMDB reviews, such as ‘actor’ and ‘performance’ (topic 2), ‘music’ and ‘soundtrack’ (topic 7)	**Overlapping themes:** some words in topics overlap, for instance, topic 0 and topic 1 both consist of words such as ‘film’, ‘movie’ or ‘watch,’ implying they all just relate to general viewing experience without clear distinction between topics, lacking clarity. Such overlap also affects the quality of the overall model.
Well-Distributed topics: BERT-topic model identifies diverse aspects of these reviews, such as acting (topic 2), director (topic 4), music (topic 6) and recommendations (topic 7).	**Ambiguous words:** some of the words in topics are ambiguous and lack interpretability, generally low-weighted too. For instance, topic 9 consists of terms such as ‘mountain,’ ‘kirk,’ or ‘rescue,’ such words are harder to semantically group together or group under a cohesive theme.

Overall, this model performs reasonably well, topics are interpretable at a moderate quality, they are usually meaningful, although there is a lot of room for improvement in reducing overlap and ambiguity. Some methods to increase coherence scores include fine-tuning the BERT-topic’s parameters and preprocessing steps.

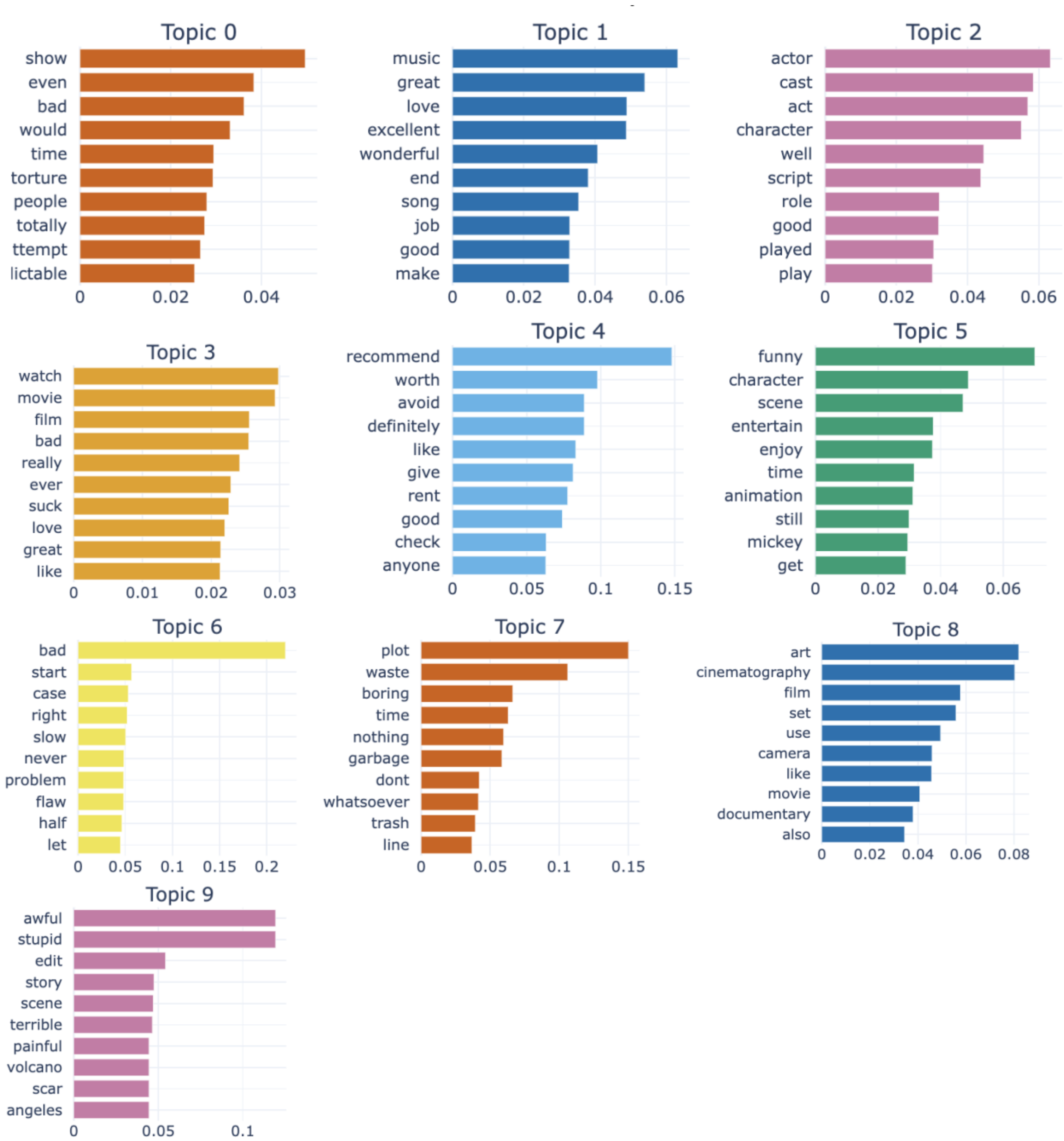
4.2 BERT-Topic with K-Means

Table 4.2a
Review and interpretation of topic and topic words

Topic	Top 10 Words	Top-Word interpretation	Overall Interpretation
0	show, even, bad, would, time, torture, people, totally, attempt, predictable	**Show:** topic likely discusses overall viewing experience or review of television series.	This review emphasises a negative viewing experience. Words such as 'bad' and 'torture' highlight an unpleasant experience.
1	music, great, love, excellent, wonderful, end, song, job, good, make	**Music**: topic focuses on soundtrack	Focus on a positive review on the film's music and soundtrack. Positive descriptors such as 'great,' 'love,' 'excellent,' and 'wonderful' heavily emphasise an appreciation for the music/soundtracks in the film
2	actor, cast, act, character, well, script, role, good, played, play	**Actor:** topic centres on actors' performance	This topic discusses a positive reflection of the overall performance and execution of the movie.
3	watch, movie, film, bad, really, ever, such, love, great, like	**Watch:** this alone suggests viewing endorsements, though other topic words, of lower weight, suggest a mixed opinion.	A mixture of positive and negative reviews with contradicting terms such as 'bad' and 'great.' It is unclear if this review is positive or negative.
4	recommend, worth, avoid, definitely, like, give, rent, good, check, anyone	**Recommend:** strong endorsement to watch the film	Focus on recommendations, highlighting a positive review and viewing experience, although 'avoid' implies some disapproval.
5	funny, character, scene, entertain, enjoy, tine, animation, still, mickey, get	**Funny:** topic likely reflects comedic appreciation of a movie	A transparent review reflecting a comedic, animated film generally perceived as funny and entertaining. A highly positive review.
6	bad, start, case, right, slow, never, problem, flaw, half, let	**Bad:** topic will criticise the film	This review criticises the pace of the film, indicating dissatisfaction towards the structure and execution of the film.
7	plot, waste, boring, time, nothing, garbage, dont, whatsoever, trash, line	**Plot:** this topic will focus on the storyline and structure of a movie.	Critique of plotlines, negative terms such as 'boring' and 'garbage' reflect disappointment in the structure and storytelling, the viewer strongly dislikes the movie, perhaps considering it a 'waste' of 'time.'
8	art, cinematography, film, set, use, camera, like, movie, documentary, also	**Art:** topic 8 will address visuals and cinematography.	This reviewer appreciated the visuals in the film, particularly the 'art' and 'cinematography.'
9	awful, stupid, edit, story, scene, terrible, painful, volcano, scar, angeles	**Awful:** this topic likely reflects a harsh criticism of a film	A harsh critique, words like 'awful,' 'terrible,' and 'painful' imply extreme dissatisfaction with the movie.

Figure 4

Visual representation of top 10 words per topic with word weight



4.2b <u>Sentiment discussion using Figure 4</u>

<u>Positive Sentiment</u> (Topics 1, 2, 4, and 5)

These bar charts consist of strongly weighted positive descriptors such as ‘great,’ ‘excellent,’ ‘recommend,’ and ‘funny.’ These topics highlighted viewers’ thorough enjoyment and appreciation for the film, especially in concepts regarding the actors’ performance, the music/soundtrack and just overall entertainment aspects.

Negative Sentiment (Topics 0, 3, 6, 7 and 9)
These charts focus on strongly weighted negative descriptors including, 'bad,' 'suck,' 'waste,' 'boring,' and 'awful,' 'predictable,' or 'torture.' Being heavily weighted words, it highlights viewers' extreme dissatisfaction, especially in aspects such as the movie plot, pace of the film or even the movie overall.

Neutral Sentiment (Topic 8)
No clear positive or negative descriptor is captured to reflect if this topic signifies a positive or negative review, seems that this review consists of a general depiction of a movie/documentary.

4.2c Quality Assessment

For this model, the coherence score is **0.420**, a moderate score but slightly lower than that of the BERT-topic model. Given the quality of both models are similar, there is similarity between model strengths and limitations:

Strengths of BERT-Topic with K-Means	Weakness of BERT-Topic with K-Means
Reasonable interpretability: many topics identified by this model are meaningful and understandable, suggesting it extracted some useful insights into IMDb reviews.	**Moderate coherence:** this score is not high enough to determine all topics are entirely meaningful, some terms may be unrelated or loosely related, effecting the quality of the overall model.
Coverage of topics: various themes were identified, such as music and soundtrack (topic 1), acting (topic 2) or humour and entertainment (topic 5); all of which are very relevant to common focus areas in IMDb reviews.	**Ambiguous words:** some words are too generic and vague, for instance, in Topic 9, words like 'volcano' and 'scar' lack any interpretable meaning, making it harder to draw useful insights into the sentiment of reviews.
Balance of sentiments: this model identified both positive and negative feedback from reviews, such balance adds depth and quality to the analysis	

Like BERT-topic, this model performed moderately, with diverse and interpretable topics, relevant to IMDb reviews and somewhat distinct, although improvements can be made for better cohesivity, particularly by addressing ambiguous terms.

4.3 Non-Negative Matrix Foundation

Table 4.3a
Review and interpretation of topic and topic words

Topic	Top 10 Words	Top-Word Interpretation	Interpretation
1	bad, act, make, series, film, write, thought, say, review, think	**Bad:** this topic likely reflects a review that critiques a movie	Strong dissatisfaction with a film, viewer thinks of it as poorly executed, particularly issues in acting and script.
2	like, movie, way, fact, film, think, hate, woman, people, attempt	**Like:** topic 1 addresses subjective opinions and cultural depictions	Mixed opinions on a film, sheds light on societal aspects, sentiment is not clear though.

3	good, actor, make, quite, job, act, cast, value, cinematography, place	**Good:** this topic will highly praise a film	Appreciation to actor, cast and cinematography, an overall positive review
4	watch, easy, joy, predictable, thing, that's, terrible, taped, love, air	**Watch:** this topic likely centres on the viewers' watching experience	Discusses the entertainment value of a film, it balances pros (easy, joy) and cons (predictable, terrible)
5	recommend, highly, saw, definitely, friend, im, fan, cinema, short, giallo	**Recommend:** topic 4 reflects a viewer's endorsement and encouragement to watch a film	Highlights a viewers' thorough enjoyment of a film, emphasising a personal endorsement and a fan of the cinema.
6	scene, plot, character, real, act, action, line, little, place, strong	**Scene:** topic 5 evaluates pivotal scenes	A positive review showing appreciation for action scenes
7	time, dont, waste, worth, long, think, money, hour, disliked, enjoy	**Time:** reflects the viewer's opinion on time spent on watching a movie	Overall assesses the film negatively, considering it not worthwhile and a waste time and money.
8	great, cast, love, director, saw, actor, end, original, disappointment, film	**Great:** likely focuses on positive aspects of a film	Viewer shows appreciation for the cast performance, directors work and the film overall, although 'disappointment' confuses what sentiment this review is.
9	really, funny, character, didn't, make, work, im, care, camera, create	**Really:** topic 8 likely emphasises positive traits of a film	Balanced view, positivity shown towards a character, but critiques aspects of the film, highlighting areas for improvement.
10	look, awful, story, end, make, cheap, advise, create, script, right	**Look:** this topic likely reflects critique in visuals	Review focuses on a viewer's disappointment in visuals and storyline, criticising it and expressing frustration over low-quality aspects.

Figure 5

Visual representation of top 10 words per topic with word weight


Topic 1
bad
act
make
series
film
write
thought
say
review
think
0.0 0.5 1.0 1.5 2.0
Word Weight
Topic 2
like
movie
way
fact
film
think
hate
woman
people
attempt
0.0 0.5 1.0 1.5 2.0
Word Weight
Topic 3
good
actor
make
quite
job
act
cast
value
cinematography
place
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
Word Weight
Topic 4
watch
easy
joy
predictable
thing
thats
terrible
taped
love
air
0.00 0.25 0.50 0.75 1.00 1.25 1.50
Word Weight
Topic 5
recommend
highly
saw
definitely
friend
im
fan
cinema
short
giallo
0.00 0.25 0.50 0.75 1.00 1.25 1.50
Word Weight
Topic 6
scene
plot
character
real
act
action
line
little
place
strong
0.0 0.2 0.4 0.6 0.8 1.0
Word Weight
Topic 7
time
dont
waste
worth
long
think
money
hour
disliked
enjoy
0.0 0.2 0.4 0.6 0.8 1.0
Word Weight
Topic 8
great
cast
love
director
saw
actor
end
original
disappointment
film
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Word Weight
Topic 9
really
funny
character
didnt
make
work
im
care
camera
create
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
Word Weight
Topic 10
look
awful
story
end
make
cheap
advise
create
script
right
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
Word Weight

4.3b **Sentiment Discussion using Figure 5**

Positive Sentiments: (Topics 3, 5 and 8)
These topic charts contain very strongly weighted positive descriptors like ‘good,’ ‘recommend,’ ‘highly,’ ‘great,’ and ‘love;’ indicating the viewers enjoyed their film, especially in aspects such as acting, plotlines and overall entertainment.

Negative Sentiments: (Topics 1, 7, 10)
These charts reveal strong weight distribution in negative descriptors, including ‘bad,’ ‘waste,’ ‘awful,’ ‘cheap,’ ‘disliked,’ and more. These highlight the viewers’ disappointment and dissatisfaction in acting (topic 1), time and money spent (topic 7) or script quality (topic 10).

Neutral/Mixed Sentiments: (Topics 2, 4, 6)
These topics are not distinctly positive or negative. Topic 2 may discuss general opinions about film productions and personal societal opinions, while topic 6 centres on specific elements of the movie such as character and plot structure, but without a clear evaluative tone. Topics 4 and 9 talk in an evaluative tone, but consists of both positive and negative features, praising the movie, but identifying some aspects as predictable and terrible (topic 4) or simply did not work (topic 9).

4.3c **Quality Assessment**

For the NMF Model, coherence equated to 0.553, outperforming both BERT-topic and K-Means. Nevertheless, this score is still moderate, suggesting the model produces interpretable topics, but there is still noise in the data preventing an increase in cohesivity.

Strengths of NMF	**Limitations of NMF**
Interpretability: topics are fairly interpretable, better than that of BERT-topic and K-Means, given a higher coherence score; topics are better at providing more meaningful insights into IMDb reviews.	**Overlapping topics**: words such as ‘act’ (1, 3, 6) and ‘film’ (topic 1, 2, 8) are present in many topics, which blurs distinction between them. Lack of distinction between topics reduces the quality of the model.
Distinctiveness and Relevance: each topic identifies interpretable and specific themes, with minimal overlap. For example, topic 5 centres on positive recommendations, using words such as ‘recommend,’ and ‘highly,’ while topic 7 focuses on viewer’s frustration capturing words like ‘time,’ ‘waste,’ and ‘money.’ Such distinctiveness increases the quality of the model.	**Mixed sentiments:** some topics (2, 4, 6) lacked clarity in distinct sentiment, it was harder to classify if these topics refer to a positive or negative review. A lack of clarity like this reduces the model’s useful insights, hindering its quality.
Diverse range of words: various descriptors identified, conveying the depth and opinions of topics. For example, topic 4 highlights specific aspects of movies (plot, acting, cinematography). Such diversity increases the quality of the NMF model.	

The NMF model performed the best across all three algorithms, nevertheless, it has still performed moderately. This model may perform substantially better if issues such as overlapping topics and unclear sentiments are addressed.

References

Arushiprakash. (n.d.). *MachineLearning/BERT Word Embeddings.ipynb at main · arushiprakash/MachineLearning*. GitHub. https://github.com/arushiprakash/MachineLearning/blob/main/BERT%20Word%20Embeddings.ipynb

Choubey, V. (2021, December 15). Topic modelling using NMF - Voice Tech Podcast - medium. *Medium*. https://medium.com/voice-tech-podcast/topic-modelling-using-nmf-2f510d962b6e

Politi, M. (2022, July 6). Feature Extraction with BERT for Text Classification. *Medium*. https://towardsdatascience.com/feature-extraction-with-bert-for-text-classification-533dde44dc2f

Singh, S. (2024, December 8). *How to Get Started with NLP – 6 Unique Methods to Perform Tokenization*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/

Thabresh. (2023, June 1). *BERT model for text embeddings*. Kaggle. https://www.kaggle.com/code/thabresh/bert-model-for-text-embeddings

Navlani, A. (2024, August 11). *Python logistic regression tutorial with Sklearn & Scikit*. DataCamp. https://www.datacamp.com/tutorial/understanding-logistic-regression-python