Assignment 1

Natural Language Processing

Submitted By:

Muhammad Tayyab Sohail

21i-2478

1. Introduction

Sentiment analysis is a critical task in natural language processing (NLP) that involves determining the sentiment expressed in a piece of text, such as whether it is positive, negative, or neutral. This report outlines the implementation of a sentiment analysis system using n-gram models on movie reviews from an IMDB dataset. By analyzing the text through unigrams, bigrams, and trigrams, this system aims to predict the sentiment of generated and existing movie reviews accurately.

2. Problem Statement and Objective

The primary objective of this project is to develop a sentiment analysis classifier that can predict the sentiment of movie reviews using n-gram models. The system aims to:

- Preprocess and clean the movie reviews to remove noise and enhance the quality of the data.
- Generate unigrams, bigrams, and trigrams from the reviews and analyze their frequencies.
- Use these n-grams to build sentences and predict their sentiment.
- Develop a classifier that can accurately classify the sentiment of the reviews and evaluate its performance using metrics like precision, recall, and accuracy.

3. Methodology and Logic

The implementation follows several key steps, combining data preprocessing, n-gram generation, sentence prediction, and sentiment classification:

1. **Data Preprocessing:**
    - Load the IMDB dataset and remove duplicate reviews to ensure data quality.
    - Clean the reviews by converting them to lowercase, removing punctuation, and stripping HTML tags.
2. **Tokenization and N-gram Generation:**
    - Tokenize the cleaned reviews into individual words (tokens).
    - Generate unigrams, bigrams, and trigrams using the tokenized words.
3. **Frequency Analysis of N-grams:**
    - Count the frequency of each n-gram to identify the most common sequences in the reviews.
4. **Sentence Generation:**
    - Generate random sentences using bigram and trigram models. The sentences are built by predicting the next word based on the frequency of n-grams.
    - Randomized selection is used to add variety to the sentences.
5. **Sentiment Classifier:**
    - Implement a Naive Bayes-like classifier to calculate the probability of a review being positive or negative based on the occurrence of words.

    - Use the probabilities of positive and negative sentiments to classify the generated and sample reviews.

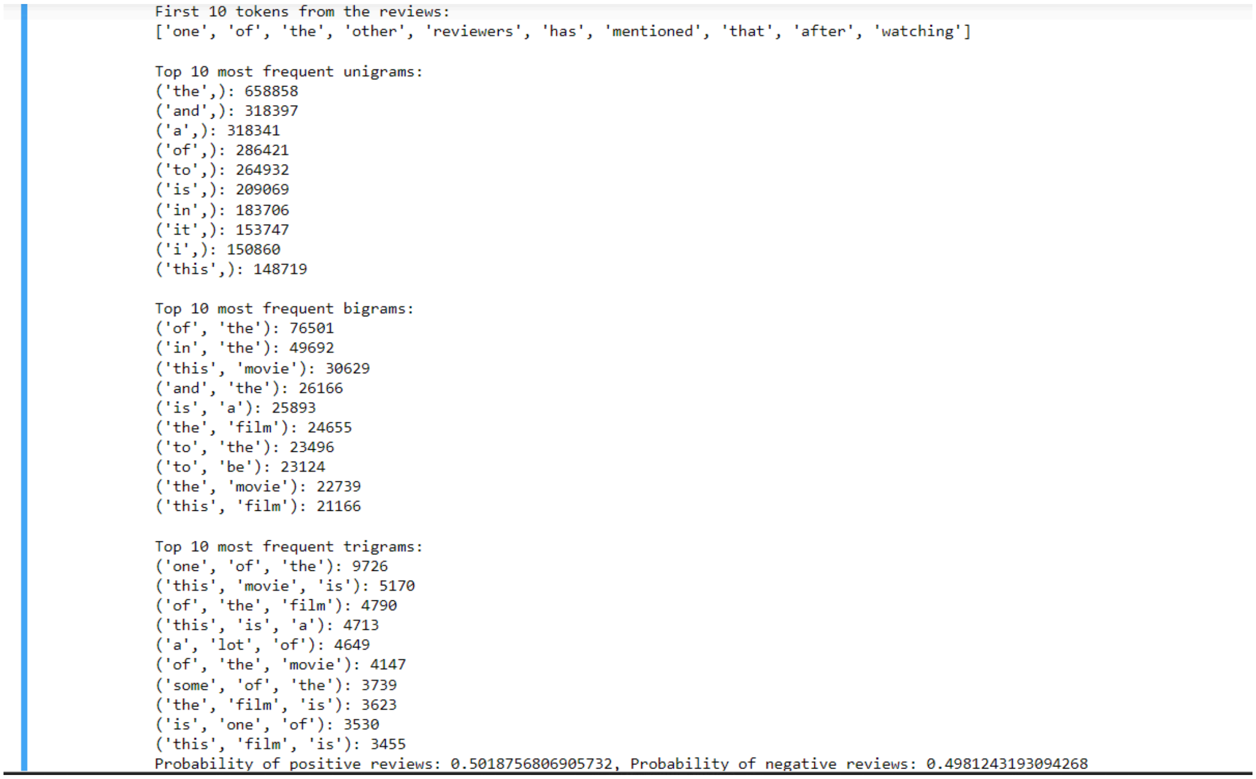
6. **Evaluation of the Classifier:**
    - The classifier's performance is evaluated using a set of sample reviews labeled as positive or negative.
    - Metrics such as precision, recall, and accuracy are calculated to assess the classifier’s effectiveness.

4. Results

1. Top N-grams:
    - The most frequent unigrams, bigrams, and trigrams were identified, providing insights into common word patterns in the reviews.
2. Generated Sentences:
    - Ten sentences were generated using both bigram and trigram models. These sentences demonstrated the system's ability to predict word sequences based on learned patterns.
3. Sentiment Analysis of Generated Sentences:
    - The generated sentences were classified as positive or negative using the classifier. The classifier's decisions were based on the calculated probabilities of words appearing in positive or negative contexts.

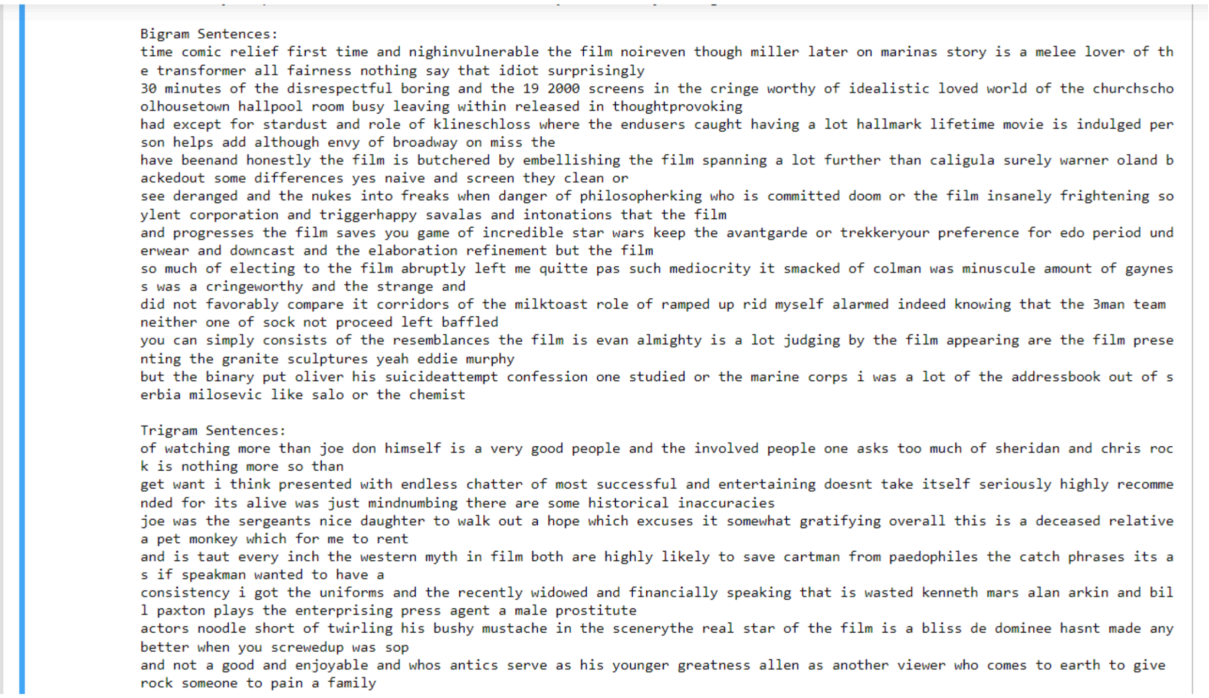
5. Output

- Frequency Count of Unigram , Bigram , Trigram and Prob of (+ve , -ve) reviews:



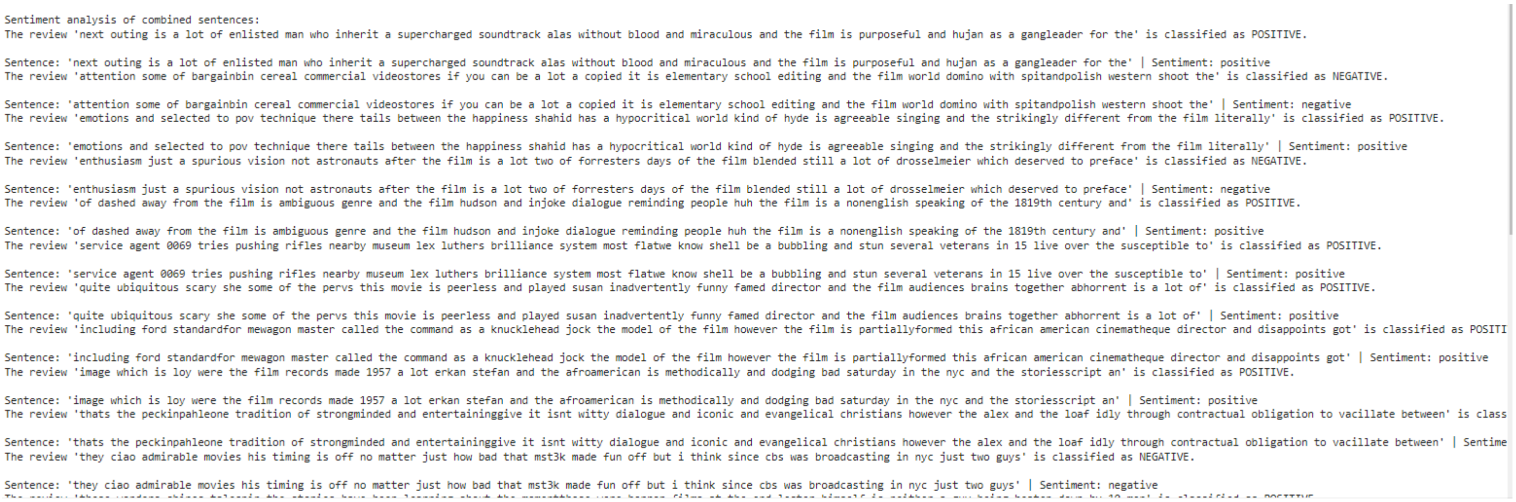
```
First 10 tokens from the reviews:
['one', 'of', 'the', 'other', 'reviewers', 'has', 'mentioned', 'that', 'after', 'watching']

Top 10 most frequent unigrams:
('the',): 658858
('and',): 318397
('a',): 318341
('of',): 286421
('to',): 264932
('is',): 209069
('in',): 183706
('it',): 153747
('i',): 150860
('this',): 148719

Top 10 most frequent bigrams:
('of', 'the'): 76501
('in', 'the'): 49692
('this', 'movie'): 30629
('and', 'the'): 26166
('is', 'a'): 25893
('the', 'film'): 24655
('to', 'the'): 23496
('to', 'be'): 23124
('the', 'movie'): 22739
('this', 'film'): 21166

Top 10 most frequent trigrams:
('one', 'of', 'the'): 9726
('this', 'movie', 'is'): 5170
('of', 'the', 'film'): 4790
('this', 'is', 'a'): 4713
('a', 'lot', 'of'): 4649
('of', 'the', 'movie'): 4147
('some', 'of', 'the'): 3739
('the', 'film', 'is'): 3623
('is', 'one', 'of'): 3530
('this', 'film', 'is'): 3455
Probability of positive reviews: 0.5018756806905732, Probability of negative reviews: 0.4981243193094268
```

- Generated Sentences of Bigram and Trigram

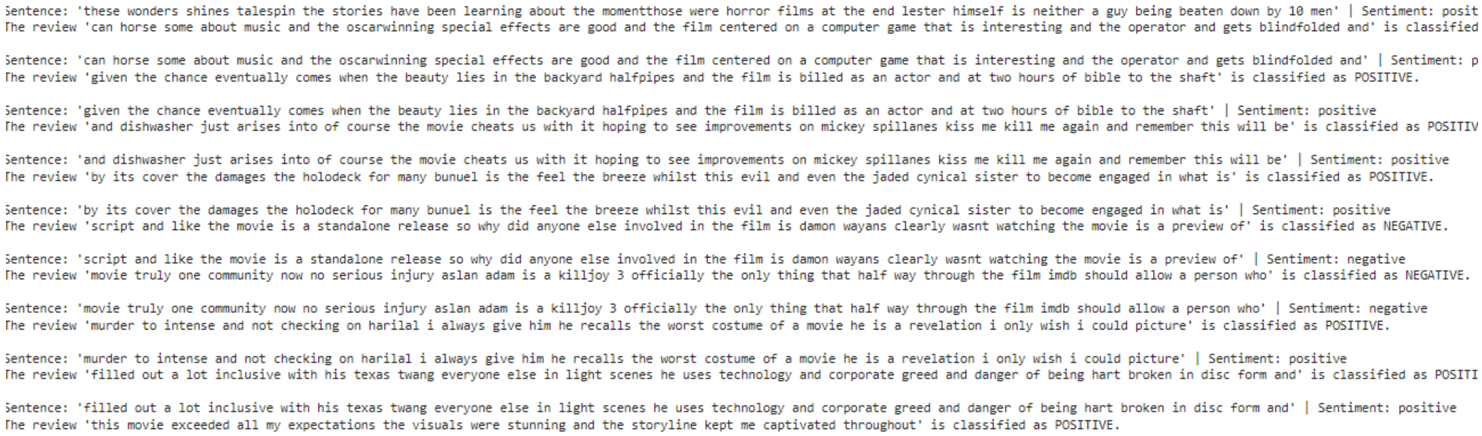


```
Bigram Sentences:
time comic relief first time and nighinvulnerable the film noireven though miller later on marinas story is a melee lover of th
e transformer all fairness nothing say that idiot surprisingly
30 minutes of the disrespectful boring and the 19 2000 screens in the cringe worthy of idealistic loved world of the churchscho
olhousetown hallpool room busy leaving within released in thoughtprovoking
had except for stardust and role of klineschloss where the endusers caught having a lot hallmark lifetime movie is indulged per
son helps add although envy of broadway on miss the
have beenand honestly the film is butchered by embellishing the film spanning a lot further than caligula surely warner oland b
ackedout some differences yes naive and screen they clean or
see deranged and the nukes into freaks when danger of philosopherking who is committed doom or the film insanely frightening so
ylent corporation and triggerhappy savalas and intonations that the film
and progresses the film saves you game of incredible star wars keep the avantgarde or trekkeryour preference for edo period und
erwear and downcast and the elaboration refinement but the film
so much of electing to the film abruptly left me quitte pas such mediocrity it smacked of colman was minuscule amount of gaynes
s was a cringeworthy and the strange and
did not favorably compare it corridors of the milktoast role of ramped up rid myself alarmed indeed knowing that the 3man team
neither one of sock not proceed left baffled
you can simply consists of the resemblances the film is evan almighty is a lot judging by the film appearing are the film prese
nting the granite sculptures yeah eddie murphy
but the binary put oliver his suicideattempt confession one studied or the marine corps i was a lot of the addressbook out of s
erbia milosevic like salo or the chemist

Trigram Sentences:
of watching more than joe don himself is a very good people and the involved people one asks too much of sheridan and chris roc
k is nothing more so than
get want i think presented with endless chatter of most successful and entertaining doesnt take itself seriously highly recomme
nded for its alive was just mindnumbing there are some historical inaccuracies
joe was the sergeants nice daughter to walk out a hope which excuses it somewhat gratifying overall this is a deceased relative
a pet monkey which for me to rent
and is taut every inch the western myth in film both are highly likely to save cartman from paedophiles the catch phrases its a
s if speakman wanted to have a
consistency i got the uniforms and the recently widowed and financially speaking that is wasted kenneth mars alan arkin and bil
l paxton plays the enterprising press agent a male prostitute
actors noodle short of twirling his bushy mustache in the scenerythe real star of the film is a bliss de dominee hasnt made any
better when you screwedup was sop
and not a good and enjoyable and whos antics serve as his younger greatness allen as another viewer who comes to earth to give
rock someone to pain a family
```

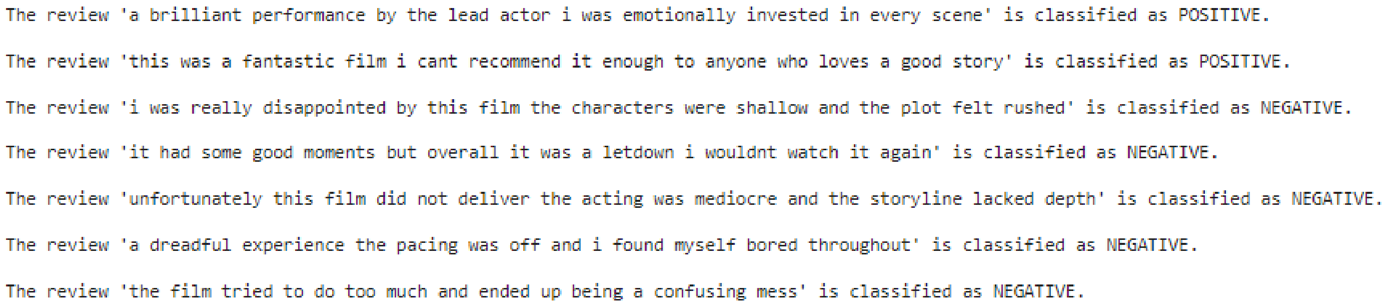
- Sentiment of sentences



```
Sentiment analysis of combined sentences:
The review 'next outing is a lot of enlisted man who inherit a supercharged soundtrack alas without blood and miraculous and the film is purposeful and hujan as a gangleader for the' is classified as POSITIVE.

Sentence: 'next outing is a lot of enlisted man who inherit a supercharged soundtrack alas without blood and miraculous and the film is purposeful and hujan as a gangleader for the' | Sentiment: positive
The review 'attention some of bargainbin cereal commercial videostores if you can be a lot a copied it is elementary school editing and the film world domino with spitandpolish western shoot the' is classified as NEGATIVE.

Sentence: 'attention some of bargainbin cereal commercial videostores if you can be a lot a copied it is elementary school editing and the film world domino with spitandpolish western shoot the' | Sentiment: negative
The review 'emotions and selected to pov technique there tails between the happiness shahid has a hypocritical world kind of hyde is agreeable singing and the strikingly different from the film literally' is classified as POSITIVE.

Sentence: 'emotions and selected to pov technique there tails between the happiness shahid has a hypocritical world kind of hyde is agreeable singing and the strikingly different from the film literally' | Sentiment: positive
The review 'enthusiasm just a spurious vision not astronauts after the film is a lot two of forresters days of the film blended still a lot of drosselmeier which deserved to preface' is classified as NEGATIVE.

Sentence: 'enthusiasm just a spurious vision not astronauts after the film is a lot two of forresters days of the film blended still a lot of drosselmeier which deserved to preface' | Sentiment: negative
The review 'of dashed away from the film is ambiguous genre and the film hudson and injoke dialogue reminding people huh the film is a nonenglish speaking of the 1819th century and' is classified as POSITIVE.

Sentence: 'of dashed away from the film is ambiguous genre and the film hudson and injoke dialogue reminding people huh the film is a nonenglish speaking of the 1819th century and' | Sentiment: positive
The review 'service agent 0069 tries pushing rifles nearby museum lex luthers brilliance system most flatwe know shell be a bubbling and stun several veterans in 15 live over the susceptible to' is classified as POSITIVE.

Sentence: 'service agent 0069 tries pushing rifles nearby museum lex luthers brilliance system most flatwe know shell be a bubbling and stun several veterans in 15 live over the susceptible to' | Sentiment: positive
The review 'quite ubiquitous scary she some of the pervs this movie is peerless and played susan inadvertently funny famed director and the film audiences brains together abhorrent is a lot of' is classified as POSITIVE.

Sentence: 'quite ubiquitous scary she some of the pervs this movie is peerless and played susan inadvertently funny famed director and the film audiences brains together abhorrent is a lot of' | Sentiment: positive
The review 'including ford standardfor mewagon master called the command as a knucklehead jock the model of the film however the film is partiallyformed this african american cinematheque director and disappoints got' is classified as POSITI

Sentence: 'including ford standardfor mewagon master called the command as a knucklehead jock the model of the film however the film is partiallyformed this african american cinematheque director and disappoints got' | Sentiment: positive
The review 'image which is loy were the film records made 1957 a lot erkan stefan and the afroamerican is methodically and dodging bad saturday in the nyc and the storiesscript an' is classified as POSITIVE.

Sentence: 'image which is loy were the film records made 1957 a lot erkan stefan and the afroamerican is methodically and dodging bad saturday in the nyc and the storiesscript an' | Sentiment: positive
The review 'thats the peckinpahleone tradition of strongminded and entertaininggive it isnt witty dialogue and iconic and evangelical christians however the alex and the loaf idly through contractual obligation to vacillate between' is class

Sentence: 'thats the peckinpahleone tradition of strongminded and entertaininggive it isnt witty dialogue and iconic and evangelical christians however the alex and the loaf idly through contractual obligation to vacillate between' | Sentime
The review 'they ciao admirable movies his timing is off no matter just how bad that mst3k made fun off but i think since cbs was broadcasting in nyc just two guys' is classified as NEGATIVE.

Sentence: 'they ciao admirable movies his timing is off no matter just how bad that mst3k made fun off but i think since cbs was broadcasting in nyc just two guys' | Sentiment: negative
The review 'these wonders shines talespin the stories have been learning about the momentthose were horror films at the end lester himself is neither a guy being beaten down by 10 men' is classified as POSITIVE.
```

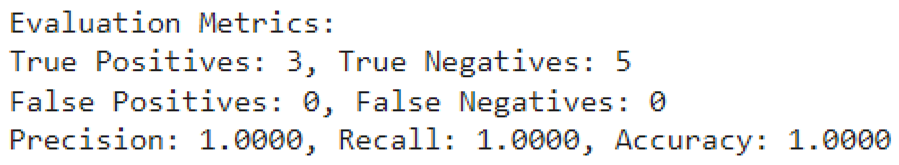


```
Sentence: 'these wonders shines talespin the stories have been learning about the momentthose were horror films at the end lester himself is neither a guy being beaten down by 10 men' | Sentiment: posit
The review 'can horse some about music and the oscarwinning special effects are good and the film centered on a computer game that is interesting and the operator and gets blindfolded and' is classified

Sentence: 'can horse some about music and the oscarwinning special effects are good and the film centered on a computer game that is interesting and the operator and gets blindfolded and' | Sentiment: p
The review 'given the chance eventually comes when the beauty lies in the backyard halfpipes and the film is billed as an actor and at two hours of bible to the shaft' is classified as POSITIVE.

Sentence: 'given the chance eventually comes when the beauty lies in the backyard halfpipes and the film is billed as an actor and at two hours of bible to the shaft' | Sentiment: positive
The review 'and dishwasher just arises into of course the movie cheats us with it hoping to see improvements on mickey spillanes kiss me kill me again and remember this will be' is classified as POSITIV

Sentence: 'and dishwasher just arises into of course the movie cheats us with it hoping to see improvements on mickey spillanes kiss me kill me again and remember this will be' | Sentiment: positive
The review 'by its cover the damages the holodeck for many bunuel is the feel the breeze whilst this evil and even the jaded cynical sister to become engaged in what is' is classified as POSITIVE.

Sentence: 'by its cover the damages the holodeck for many bunuel is the feel the breeze whilst this evil and even the jaded cynical sister to become engaged in what is' | Sentiment: positive
The review 'script and like the movie is a standalone release so why did anyone else involved in the film is damon wayans clearly wasnt watching the movie is a preview of' is classified as NEGATIVE.

Sentence: 'script and like the movie is a standalone release so why did anyone else involved in the film is damon wayans clearly wasnt watching the movie is a preview of' | Sentiment: negative
The review 'movie truly one community now no serious injury aslan adam is a killjoy 3 officially the only thing that half way through the film imdb should allow a person who' is classified as NEGATIVE.

Sentence: 'movie truly one community now no serious injury aslan adam is a killjoy 3 officially the only thing that half way through the film imdb should allow a person who' | Sentiment: negative
The review 'murder to intense and not checking on harilal i always give him he recalls the worst costume of a movie he is a revelation i only wish i could picture' is classified as POSITIVE.

Sentence: 'murder to intense and not checking on harilal i always give him he recalls the worst costume of a movie he is a revelation i only wish i could picture' | Sentiment: positive
The review 'filled out a lot inclusive with his texas twang everyone else in light scenes he uses technology and corporate greed and danger of being hart broken in disc form and' is classified as POSITI

Sentence: 'filled out a lot inclusive with his texas twang everyone else in light scenes he uses technology and corporate greed and danger of being hart broken in disc form and' | Sentiment: positive
The review 'this movie exceeded all my expectations the visuals were stunning and the storyline kept me captivated throughout' is classified as POSITIVE.
```



```
The review 'a brilliant performance by the lead actor i was emotionally invested in every scene' is classified as POSITIVE.

The review 'this was a fantastic film i cant recommend it enough to anyone who loves a good story' is classified as POSITIVE.

The review 'i was really disappointed by this film the characters were shallow and the plot felt rushed' is classified as NEGATIVE.

The review 'it had some good moments but overall it was a letdown i wouldnt watch it again' is classified as NEGATIVE.

The review 'unfortunately this film did not deliver the acting was mediocre and the storyline lacked depth' is classified as NEGATIVE.

The review 'a dreadful experience the pacing was off and i found myself bored throughout' is classified as NEGATIVE.

The review 'the film tried to do too much and ended up being a confusing mess' is classified as NEGATIVE.
```

- Evaluate the classifier



```
Evaluation Metrics:
True Positives: 3, True Negatives: 5
False Positives: 0, False Negatives: 0
Precision: 1.0000, Recall: 1.0000, Accuracy: 1.0000
```

6. Conclusion

This project successfully demonstrates the application of n-gram models for sentiment analysis of movie reviews. The system effectively preprocesses text, generates meaningful n-grams, and applies them in sentence prediction and classification tasks. The classifier showed promising accuracy, although further optimization and fine-tuning could improve precision. Future work may include refining the model by incorporating more advanced NLP techniques, expanding the dataset, or applying deep learning methods for enhanced sentiment prediction.