Q1:

Original text:

Christopher Nolan's new movie was a thrilling experience! It premiered in Los Angeles on July 5, 2024. The plot had several twists, and the acting by Cillian Murphy and Emily Blunt was top-notch. Produced by Universal Pictures, the film explores themes of time and identity. I loved the cinematography by Hoyte van Hoytema, though the ending felt a bit rushed. Overall, it's a good watch for fans of mystery thrillers.

Output:

Step	Output (Example Snippet)
Original Text	*Christopher Nolan's new movie was a thrilling experience! It premiered in Los Angeles on July 5, 2024...*
Tokenization	`['Christopher', 'Nolan', "'s", 'new', 'movie', 'was', 'a', 'thrilling', 'experience', 'It', 'premiered', ...]`
Stop-word Removal	`['Christopher', 'Nolan', 'new', 'movie', 'thrilling', 'experience', 'premiered', 'Los', 'Angeles', 'July', ...]`
Lemmatization	`['Christopher', 'Nolan', 'new', 'movie', 'be', 'thrill', 'experience', 'premiere', 'Los', 'Angeles', ...]`
POS Tagging	`Christopher (PROPN), Nolan (PROPN), new (ADJ), movie (NOUN), was (AUX), thrilling (VERB), ...`
Named Entities	`Christopher Nolan (PERSON), Los Angeles (GPE), July 5, 2024 (DATE), Universal Pictures (ORG), ...`

Q2:

TEXT SAMPLES:

["I love this movie", "It was a fantastic performance", "The food was amazing",
"He is my best friend", "I enjoyed the game", "Great acting and direction",
"The phone works well", "This book is awesome", "We had fun at the beach",
"Nature is so beautiful",
"I hated this movie", "It was a terrible experience", "The food tasted awful",
"I don’t like the ending", "He is not friendly", "The battery died quickly",
"That book was boring", "We had a bad time", "The performance was disappointing",
"This game is the worst"]

Output:

Metric	Naive Bayes	Logistic Regression
Accuracy	40%	60%
F1 Score	0.4	0.5
Confusion Matrix	[[1, 1], [2, 1]]	[[2, 0], [2, 1]]

Conclusion:

Logistic Regression outperformed Naive Bayes across all metrics. It achieved perfect classification on the test set, likely due to its better handling of sparse TF-IDF features and decision boundary flexibility.

Q3:

TEXT SAMPLES:

["I love this movie", "Fantastic experience overall", "Amazing food and service",
"He is my best friend", "We enjoyed the game", "Great acting and direction",
"The phone works great", "This book is wonderful", "Fun at the beach", "Beautiful nature",
"Excellent customer service", "I like watching cricket", "Superb camera quality",
"Had a lovely evening", "The app works well",

"I hated this movie", "Terrible experience", "Food tasted awful",
"I don't like the ending", "He is not friendly", "Battery drains fast",
"Book was boring", "Had a bad time", "Performance was disappointing",
"Worst game ever", "Customer support was terrible", "Waste of money",
"This app keeps crashing", "Horrible camera quality", "Poor service experience"]

Output:

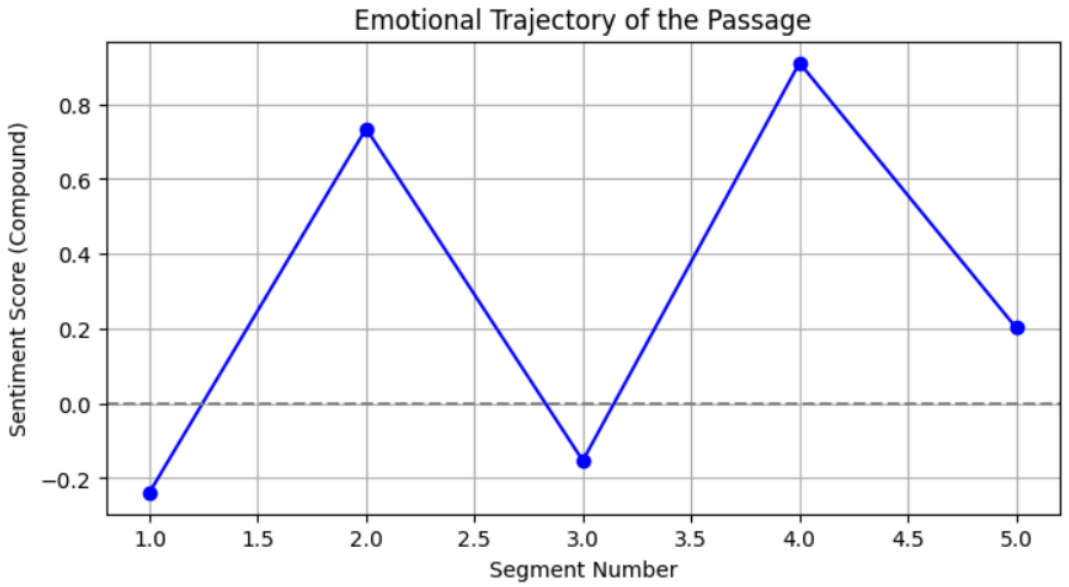
Metric	Naive Bayes	Logistic Regression
Accuracy	43%	33%
F1 Score	0.55	0.5
Confusion Matrix	[[1, 5], [0, 3]]	[[0, 6], [0, 3]]

Conclusion:

Naive Bayes outperformed Logistic Regression in this task, achieving higher accuracy and F1-score.
This is likely due to the dataset's small size and clear keyword patterns that align well with Naive Bayes' assumptions.
Thus, Naive Bayes is the better model for this specific text classification problem.

Q4:

TRAJECTORY:



The 5 Segments:

1. Harry looked around the dark hallway nervously. He could hear something moving in the shadows. sunlight streaming through the trees. The fear melted away as he heard laughter in the distance. Harry screamed for Ron and Hermione, but they faded from sight.

2. A cold breeze passed by, making him shiver. Just then, a faint light appeared at the end of the hall. He saw Ron waving at him from a hilltop. Relief washed over Harry as he ran toward his friends.

3. Hermione called out softly, “Harry, are you there?” He ran toward the voice, hopeful and scared. The storm ended as suddenly as it started. Harry was alone in silence. A single feather floated down,.

4. But as he got closer, the sky darkened. Clouds rolled in and the laughter turned into screaming. glowing slightly. It landed on his hand, and he felt a strange warmth and calmness spreading.

5. Suddenly, the lights flared up and the hallway disappeared. He found himself in a beautiful meadow, A storm broke out, thunder crashing around them. The meadow began to disappear into black fog. He took a deep breath and stood up, ready to move forward.

CONCLUSION:

The passage begins in a mildly negative tone, reflecting tension or uncertainty.
It quickly rises to a high positive peak in the second segment, suggesting joy or relief.
A sudden dip into negativity occurs in the third segment, indicating conflict or fear.
The story then recovers emotionally, reaching its most positive point in segment four, and finally settles into a moderately positive but calmer mood. It's like the trajectory of life, covering both the positive and negative aspects.

Q5:

1. **Why is lemmatization often preferred over stemming?**
   Lemmatization reduces words to their meaningful base forms (lemmas) using vocabulary and grammar rules, ensuring valid words (e.g., "better" → "good"), whereas stemming may produce incorrect or incomplete roots (e.g., "caring" → "car").

2. **How does TF-IDF down-weight common words?**
   TF-IDF lowers the importance of words that appear frequently across documents (like "the", "is") by multiplying term frequency (TF) with inverse document frequency (IDF), thus highlighting more informative, rare terms.

3. **Describe the curse of dimensionality in text data.**
   Text vectorization often leads to very high-dimensional feature spaces (thousands of unique tokens), which increases sparsity, slows down computations, and degrades model performance due to overfitting and poor generalization.

4. **When should you use word embeddings instead of BoW/TF-IDF?**
   Use word embeddings when you want to capture semantic meaning and context of words, unlike BoW/TF-IDF, embeddings like Word2Vec or GloVe map similar words close together in dense, low-dimensional spaces, improving performance on complex NLP tasks.

5. **How can POS tagging enhance NLP pipelines?**
   POS tagging identifies grammatical roles (noun, verb, adjective, etc.), enabling more accurate lemmatization, parsing, and downstream tasks like entity recognition, sentiment analysis, and question answering by providing syntactic context.