DOCUMENTATION

PROBLEM:

Next Word Prediction is solving a *sequence prediction* problem. Specifically, it learns patterns in sequences of words to predict what word is likely to come next in a sentence, based on the context of previous words.

This type of problem is useful in applications like autocomplete, where we want to suggest the next word a user might type, or in machine translation, where predicting the next word in a sequence can help generate accurate translations.

DATASET:

Description: The dataset consists of text from **Pride and Prejudice** by **Jane Austen**, formatted as an eBook. This text includes all the chapters and descriptions of characters, interactions, and storylines from the original novel.

Link: Dataset Link

Sample:

Mr. Bingley had soon made himself acquainted with all the principal people in the room: he was lively and unreserved, danced every dance, was angry that the ball closed so early, and talked of giving one himself at Netherfield. Such amiable qualities must speak for themselves. What a contrast between him and his friend! Mr. Darcy danced only once with Mrs. Hurst and once with Miss Bingley, declined being introduced to any other lady, and spent the rest of the evening in walking about the room, speaking occasionally to one of his own party. His character was decided. He was the proudest, most disagreeable man in the world, and everybody hoped that he would never come there again. Amongst the most violent against him was Mrs. Bennet, whose dislike of his general behaviour was sharpened into particular resentment by his having slighted one of her daughters.

METHOD:

LSTM (Long – Short Term Memory):

The LSTM (Long Short-Term Memory) algorithm is a type of recurrent neural network (RNN) designed to handle sequence prediction tasks, particularly where long-term dependencies are crucial. Unlike traditional RNNs, LSTMs incorporate a special memory cell and gates (input, forget, and output) that regulate the flow of information, allowing the network to retain important information over extended sequences and "forget" less useful details. This architecture makes LSTMs effective in tasks like natural language processing, speech recognition, and time-series prediction, where context from earlier parts of a sequence significantly influences outcomes. By capturing patterns over long sequences, LSTMs are ideal for applications such as text generation, translation, and next-word prediction.

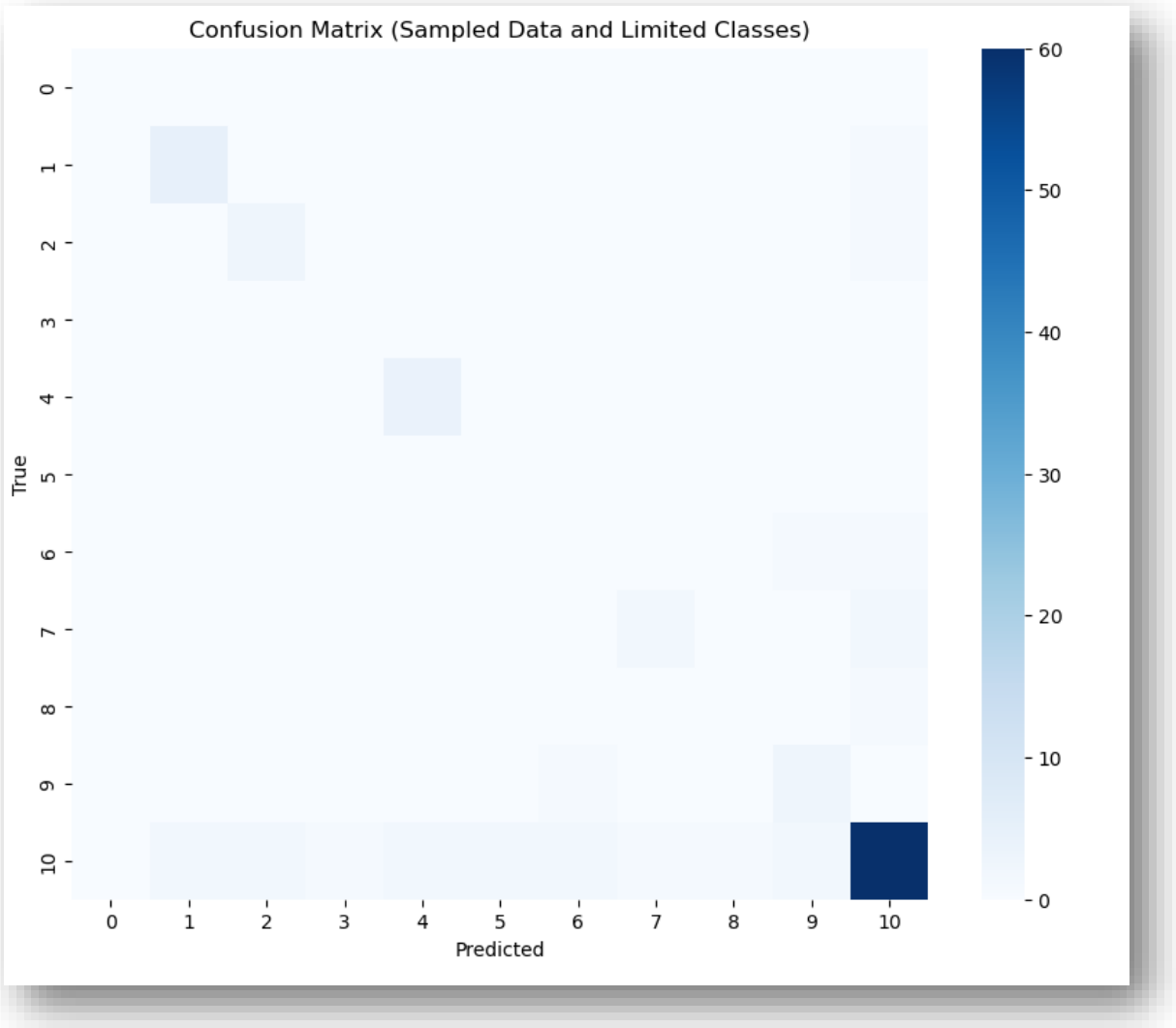
RESULTS:

Evaluation Metrics:

The notebook includes several evaluation metrics for assessing the LSTM model's performance:

1. **Accuracy**: The model is compiled to track accuracy during training.
2. **Precision and Recall**: These metrics are also tracked during training, helping to evaluate the model's predictive quality beyond accuracy alone.
3. **Confusion Matrix**: A confusion matrix is computed on a random sample of predictions, focusing on the top classes to visualize the distribution of true versus predicted labels. This matrix is displayed as a heatmap.
4. **F1 Score**: Calculated on the sampled predictions using a weighted average to account for class imbalances.

Here is the visualized version of confusion matrix for the LSTM model:



INTERPRETATION & ERROR ANALYSIS:

Confusion Matrix: The confusion matrix shows that the model predicts class 10 very well, with most correct predictions in this class. There are some misclassifications across other classes, but they are relatively low. This suggests the model performs best for class 10 but struggles somewhat with other classes, likely due to class imbalance or feature limitations.

Accuracy, Precision, Recall: Over 15 epochs, the model shows steady improvement:

- **Accuracy** rises from 5.4% to 46.97%, indicating better predictions.
- **Precision** increases from 47.29% to 88.83%, meaning fewer false positives.
- **Recall** goes up from 0.03% to 22.77%, showing progress in capturing true positives, though it remains low.

F1 Score: An F1 score of 0.79 indicates a good balance between precision and recall, meaning the model is effective at correctly identifying relevant instances while minimizing both false positives and false negatives. However, there's still room for improvement to reach even higher accuracy.

Errors: Here, I have listed the Expected output along with the actual output to see how the model is performing and where it is making wrong predictions.

Model's Output	Expected Output	Error Type
Why are you ” said	why are not you dancing	Contextual Misunderstanding Error
What is the matter ” said elizabeth “	what is the matter?” cried he	Contextual Misunderstanding Error
but her knees trembled at all	but her knees trembled under her	Subtle Semantic Differences Error
I believe you will be very	I believe you thought her rather	Length and Complexity Error
He then went on the subject	He then went away, and Miss	Subtle Semantic Differences Error
except what had passed the door	except what had particularly interested them	Contextual Misunderstanding Error
his friends may not be delayed	his friends may well rejoice in	Subtle Semantic Differences Error
My friend has been softened by	My friend has an excellent understanding	Length and Complexity Error
she could not help flattering	she could not have bestowed	Contextual Misunderstanding Error
she seems to see a	She seems perfectly happy,	Feature Sparsity Error
the feelings of her face	the feelings which were divided	Synonym and Polysemy Error
rather adds to the purpose of	rather adds to your other perfections	Contextual Misunderstanding Error
You can hardly know that i	You can hardly doubt the purport	Length and Complexity Error
I desire you will not	I desire you will stay	Subtle Semantic Differences Error

she sat intently in her	she sat down again,	Contextual Misunderstanding Error
that your sister has been	that your modesty, so far	Length and Complexity Error
rather than how much i	rather adds to your other perfections	Subtle Semantic Differences Error
but allow her to be	but allow me to assure	Synonym and Polysemy Error
Caroline neither expects nor face	Caroline neither expects nor wishes	Contextual Misunderstanding Error
hope of being to be	hope of an event which	Length and Complexity Error
elizabeth could not help smiling	Elizabeth could not oppose such	Overlapping Class Error
though the elder miss bennets	though the probability of the	Contextual Misunderstanding Error
dish was to be inferior	dish was commended first by him	Synonym and Polysemy Error
The party were in the	The party did not supply	Contextual Misunderstanding Error
Elizabeth wondered it was not	Elizabeth wondered Lady Catherine could	Contextual Misunderstanding Error
The dinner was spent in	The dinner was exceedingly handsome	Length and Complexity Error
Then pray what is your	Then pray speak aloud.	Contextual Misunderstanding Error
Let me flatter ourselves that	Let me hear what it is	Contextual Misunderstanding Error
well entertained a great air	well entertained in that room	Feature Sparsity Error
by his sister a more	by his high and imposing	Length and Complexity Error
world is a capital but	world is blinded by his	Synonym and Polysemy Error
I wish I am sure i	I wish I could call her	Subtle Semantic Differences Error
she was not very poorly	she was affectionate and pleasing	Subtle Semantic Differences Error
I have devoted a letter and	I have devoted hours and hours	Length and Complexity Error
I understand you will not	I understand, highly accomplished	Subtle Semantic Differences Error
She danced next with him and	She danced next with an officer	Length and Complexity Error
I dare say you will not	I dare say you will find him	Subtle Semantic Differences Error