Shakespearean Generative AI Lab

Bigram Language Model & Text Analysis

Fatima Abubakar

Beginner, Generative AI & Data Science

flexisaf

Statement of the Problem

Shakespeare Generative has four aims.

Objectives

- Train a beginner-level generative text model
- Use Shakespeare’s Romeo and Juliet as training data
- Analyze linguistic patterns
- Evaluate how well generated text mimics original text

Dataset

Source: Public domain text from Romeo and Juliet

- Total words: **28,959 words**
- Vocabulary size: **4,187 unique words**

Dataset created by:

- Downloading text file
- Converting to CSV format
- Storing each line as one data entry

Data Cleaning

Steps performed:

- Converted text to lowercase
- Removed punctuation
- Removed non-alphabetic characters
- Tokenized into words
- Combined into continuous corpus

Purpose:

- Ensure consistent probability calculations.

Model Approach

We built a Bigram Model:

P(next_word | current_word)

The model learns:

“love” → “is”

“is” → “a”

“a” → “smoke”

It stores word transition probabilities.

This is a simple Markov Chain.

Training Process

- Loop through all words
- Create word pairs
- Store next-word possibilities
- Sample randomly from learned transitions

Generated Sample

Example output:

“love is a smoke raised with the fume of sighs and sorrow enters where light once stood…”

Real generated output:

“love why lookst thou stay exit scene iv a week for then down their fingers capulet and spurs swits and be distraught environed with an ebook”

Evaluations

We evaluated using:

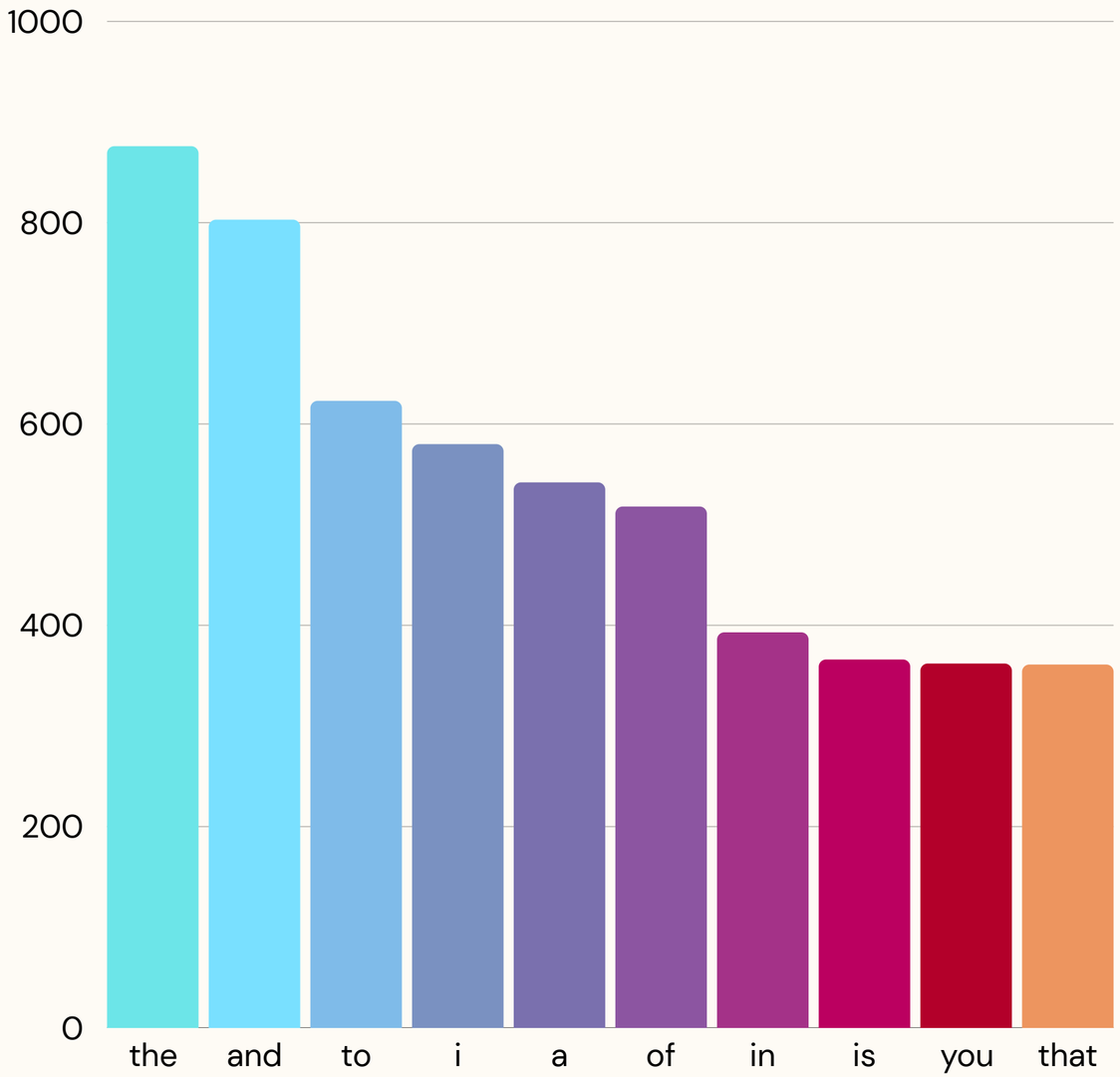
- Word frequency comparison
- Vocabulary size
- Sentence length distribution
- Visual charts

Goal:

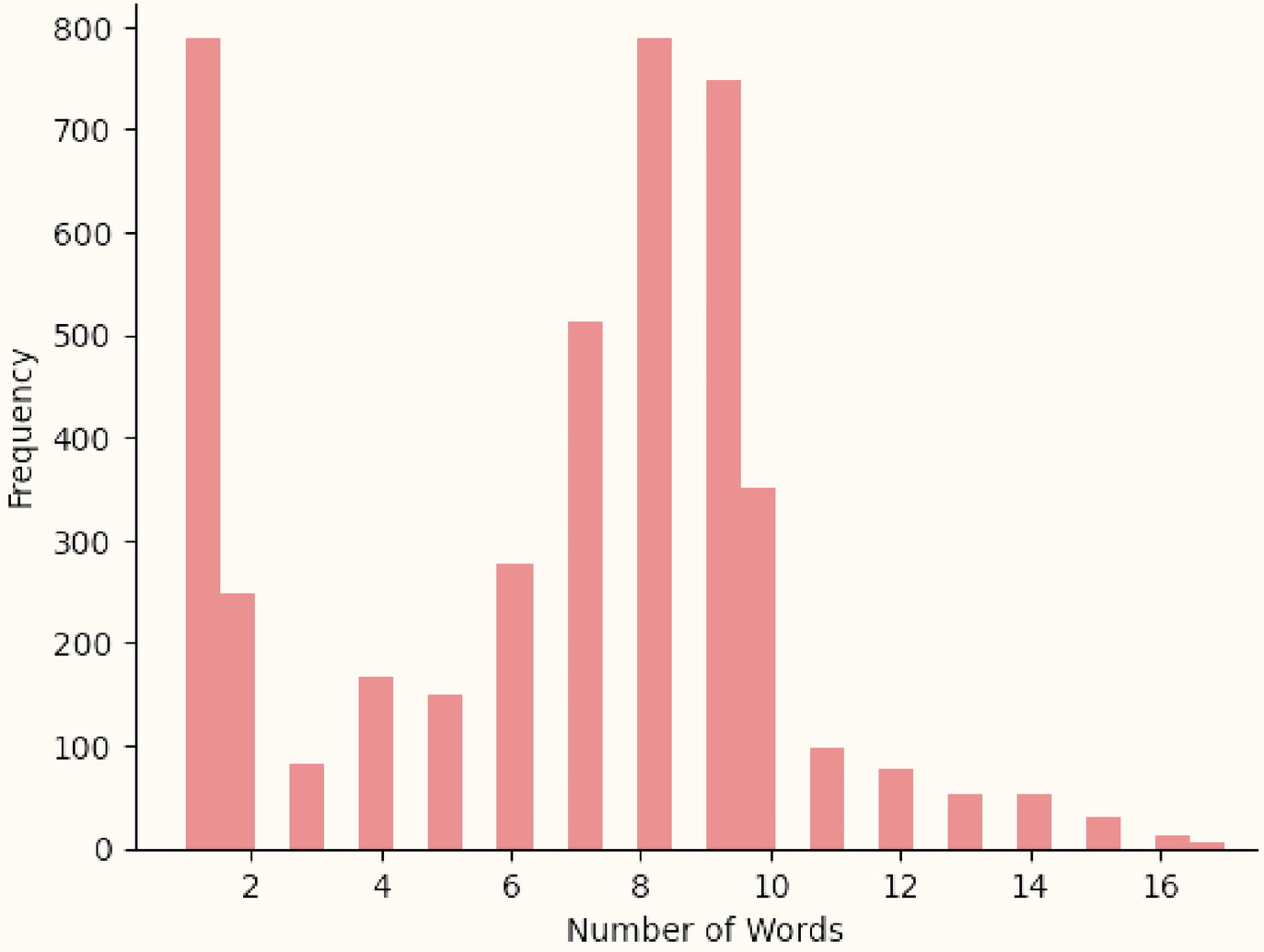
Determine how closely generated text resembles training data.

Word Frequency Chart

Observation:
High-frequency words dominate generation.



The bar chart illustrates that connective words are the most frequently used in the generation



Sentence Length Distribution

Observation:

Generated text mirrors average sentence structure but lacks semantic coherence.

It illustrates the distribution of sentence lengths in the cleaned text.

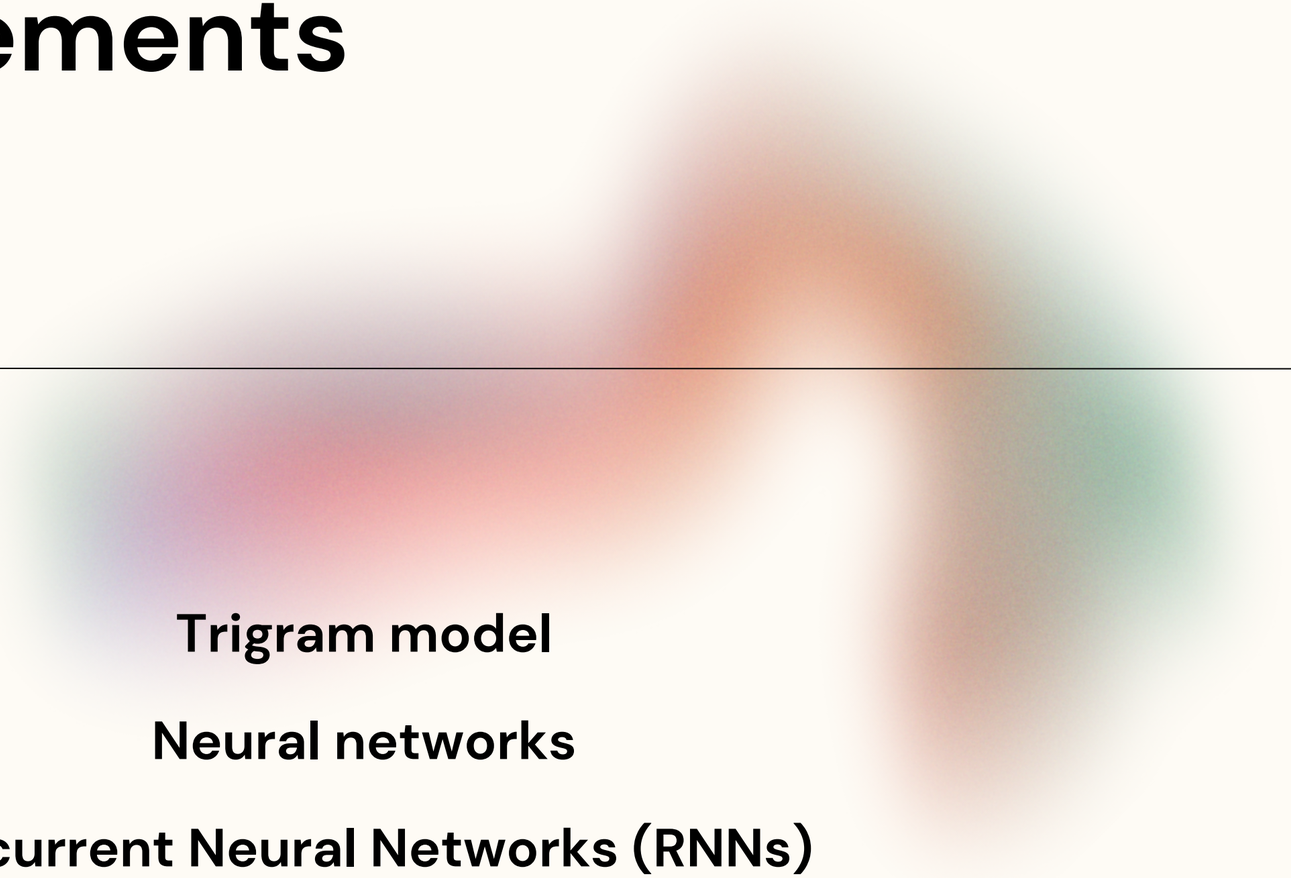
Key Insights

- The model captures local word structure.
- It cannot understand meaning.
- It lacks long-range dependencies.
- Vocabulary richness depends entirely on dataset.

Limitations

- No deep contextual understanding
- No grammar enforcement
- Limited memory (1-word window)
- Cannot reason

Future Improvements



Trigram model

Neural networks

Recurrent Neural Networks (RNNs)

Transformers (like GPT)

Conclusion

This lab demonstrates:

- How generative models learn from data
- How probability drives text generation
- The importance of data analysis in AI systems