Detailed Analysis of N-Gram Text Prediction

Mudit Gupta
Roll Number: 2024201058

March 27, 2025

1 Introduction

This report provides a comprehensive evaluation of an N-Gram model for text prediction. The analysis focuses on its accuracy, efficiency, and adaptability across various datasets, with results interpreted to assess generalization and performance trends.

2 Corpus Details and Size

Three datasets were utilized for training and evaluation:

- **General English Corpus**: A broad, diverse dataset representing standard English text.
- **Topic-Specific Dataset**: A focused corpus with domain-specific content.
- **Topic-Specific Dataset (Part 7)**: A subset featuring specialized terminology.

3 Performance Metrics

The following metrics were analyzed:

- **Total Letter Keys Typed**: Number of character keystrokes.
- **Total Tab Key Presses**: Number of Tab key completions.
- **Avg Letters per Word**: Computed as $\frac{\text{Total Letter Keys}}{\text{Total Words}}$.
- **Avg Tabs per Word**: Computed as $\frac{\text{Total Tab Keys}}{\text{Total Words}}$.

4 Experimental Results

4.1 Letter and Tab Key Analysis

Dataset	Letter Keys	Tab Keys	Avg Letters/Word	Avg Tabs/Word
General English Corpus				
n=2	297	383	0.29	1.60
n=3	276	420	0.26	1.76
n=6	263	440	0.25	1.84
n=10	263	428	0.25	1.79
Topic-Specific Dataset				
n=2	261	374	0.24	1.53
n=3	245	358	0.23	1.47
n=6	225	357	0.21	1.47
n=10	224	350	0.21	1.44
Topic-Specific Dataset (Part 7)				
n=2	233	382	0.22	1.58
n=3	226	328	0.21	1.36
n=6	208	359	0.19	1.48
n=10	209	340	0.20	1.40

Table 1: Performance metrics across datasets for different n-gram values.

4.2 Comparison Across Different N Values

Dataset	n=2	n=3	n=6	n=10
General English Corpus	1.60	1.76	1.84	1.79
Topic-Specific Dataset	1.53	1.47	1.47	1.44
Topic-Specific Part 7	1.58	1.36	1.48	1.40

Table 2: Average Tabs per Word for different n-gram values.

4.3 Video Results Summary

Video	Dataset	Model (n)	Letter Keys	Avg Letters/Word	Avg Tabs/Word
General English Test	General English Corpus	2	288	0.33	1.62
Best Model Combo	Topic-Specific Part 7	3	226	0.21	1.36

Table 3: Video results summary.

5 Findings and Discussion

5.1 Corpus Analysis

General English Corpus: The dataset showed a steady increase in tab key usage from $n = 2$ (1.60) to $n = 6$ (1.84), followed by a slight decrease at $n = 10$ (1.79). This suggests that higher n-gram values improve prediction up to a point, beyond which gains diminish, possibly due to overfitting or limited additional context.

Topic-Specific Dataset: Tab usage decreased consistently from $n = 2$ (1.53) to $n = 10$ (1.44), with stabilization between $n = 3$ and $n = 6$. This indicates that the model benefits from increased context in domain-specific text, achieving optimal efficiency at $n = 10$.

Topic-Specific Dataset (Part 7): The lowest tab usage was observed at $n = 3$ (1.36), with a slight increase at $n = 6$ (1.48) and a reduction again at $n = 10$ (1.40). This suggests that $n = 3$ captures sufficient context for this specialized subset, with marginal improvements at $n = 10$.

5.2 Model Performance Analysis

- **n=2**: Highest tab usage across datasets (1.53–1.60), indicating limited predictive capability due to insufficient context.
- **n=3**: Improved performance, with Topic-Specific Part 7 achieving the lowest tab usage (1.36), suggesting adequacy for specialized text.
- **n=6**: Mixed results—highest tab usage for General English (1.84) but stable for Topic-Specific datasets (1.47–1.48), reflecting a balance between context and complexity.
- **n=10**: Best overall performance for Topic-Specific datasets (1.44 and 1.40), though slightly less efficient for General English (1.79), indicating specialization benefits.

5.3 Video Results Analysis

The General English Corpus with $n = 2$ showed moderate tab usage (1.62) with higher avg letters per word (0.33), suggesting good generalization. The Topic-Specific Part 7 with $n = 3$ achieved the lowest tab usage (1.36), reinforcing its efficiency for domain-specific prediction.

6 Conclusion and Recommendations

- **Best Model Selection**: The $n = 3$ model excels for Topic-Specific Part 7 (1.36 Avg Tabs/Word), while $n = 10$ is optimal for broader adaptability (1.44–1.79).
- **Best Corpus Selection**: Topic-Specific Part 7 for specialized content; General English Corpus for general text.
- **Future Work**: Explore hybrid n-gram models or transformer-based methods to further reduce tab usage and enhance adaptability.