**TASK 1: Topic ModelingA graph with blue lines and numbers

AI-generated content may be incorrect.A close up of words

AI-generated content may be incorrect.**

* **Total Samples**: 5,717 (training)
* Texts vary greatly in length, from **1 word to over 11,000 words**.
* **Average length**: 317.61 words per sample.

This wide variance suggests that the dataset includes both extremely short posts and long-form discussions. Longer samples may significantly impact preprocessing decisions such as truncation, padding, or token window size in transformer-based models.

**🔍 Topic Discovery with LDA:**

* **Topic range tested**: K = 1 to 10
* **Best Coherence Score**: ~0.62 at **K = 1**
* Other strong scores: K = 6 and 7 (~0.56)
* Worst score: K = 5 (~0.41)

Despite testing multiple topics, the model determined that **a single-topic solution** offered the most coherent representation. This suggests that the dataset may revolve around a dominant theme or contains overlapping content.

**☁️ Word Cloud for Topic 1 (K=1):**

* **Top Words**: would, one, people, know, like, say, god, windows, make, time
* Indicative of open discussion, possibly forum-like or general dialogic content.
* Some noise (e.g., "maxaxaxaxaxa") points to informal writing or usernames, which may require further cleaning.

**TASK 2: Named Entity Recognition (NER) & Affective Analysis**

**NER Results**

* **Top 3 Entity Types**:
  + ORG: 28,561
  + PERSON: 28,272
  + CARDINAL: 20,314

These entity types suggest that the dataset includes a **rich amount of real-world context**, with a focus on organizations, individuals, and quantities. Such data are highly valuable for many NLP applications like relation extraction, document classification, or sentiment tracking in financial/news domains.

**Average Valence Scores (from NRC VAD Lexicon):**

* **ORG**: 1.033 → moderately positive sentiment
* **PERSON**: 0.347 → relatively neutral
* **CARDINAL**: -0.625 → mildly negative

This shows a **nuanced emotional landscape**:

* Organizations are associated with positive emotions,
* People evoke more neutral reactions,
* Numbers (CARDINALs) might appear more often in negative or critical contexts (e.g., death tolls, losses).

**Overall Sentiment (VADER Analysis):**

* Negative: 7.04%
* Neutral: 83.66%
* Positive: 9.30%
* Compound: 0.1458 (slightly positive overall)

Your dataset is **heavily neutral**, which is common in factual or journalistic text. The slight lean toward positivity in the compound score suggests a **non-hostile, potentially optimistic tone**, possibly due to positive organizational mentions.

**TASK 3: Text Classification**

**Modeling:**

* **Technique**: TF-IDF + Logistic Regression
* **Accuracy**: **94.67%**

**Performance Metrics:**

| **Class** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- |
| comp.os.ms-windows.misc | 0.93 | 1.00 | 0.96 |
| rec.motorcycles | 0.94 | 0.98 | 0.96 |
| soc.religion.christian | 0.96 | 0.96 | 0.96 |
| talk.politics.guns | 0.93 | 0.92 | 0.93 |
| talk.politics.mideast | 0.99 | 0.90 | 0.94 |
| talk.politics.misc | 0.95 | 0.87 | 0.91 |

* **Macro F1-Score**: 0.94
* **Weighted F1-Score**: 0.95

The model demonstrates **excellent generalization across all categories**, especially for tech, religion, and motorcycle classes. Slight dips in recall for the political categories may reflect **greater lexical ambiguity or topic overlap**, which is common in opinionated or cross-topic discussions.