SemEval Task 4: Narrative Similarity Track A

Exploratory Data Analysis for CS 14B

Muhammad Ahmed (510253)
, Masab Bin Imtiaz (507715)
, Muhammad Salar Abdullah (503603)
, Rana Jahanzaib Ali (467345)
Department of Computer Science, NUST

***Abstract*—This report presents the exploratory data analysis (EDA) for Track A of SemEval Task 4: Narrative Similarity. All core and advanced data visualizations are included, from text length and label distribution to similarity metrics and embedding spaces. Each plot contributor is credited simply in parentheses at the end of the figure caption.**

I. Introduction

SemEval Task 4 centers on narrative similarity: given an anchor story, the goal is to select which of two candidates (A or B) is closer in meaning. The CS 14B group (Muhammad Ahmed (510253), Masab Bin Imtiaz (507715), Muhammad Salar Abdullah (503603), Rana Jahanzaib Ali (467345)) performed EDA on the development dataset for Track A.

II. Dataset Overview

The dataset consists of triples—anchor, text A, text B. Example:

Anchor: The book follows an international organization named the Ministry for the Future...
A: The old grandmother Tina arrives in town to attend the wedding of his nephew Alberto...
B: Glenn Tyler (Elvis Presley), a childish 25-year-old from a troubled background...

III. EDA Results and Plots

A. Text Length Distribution

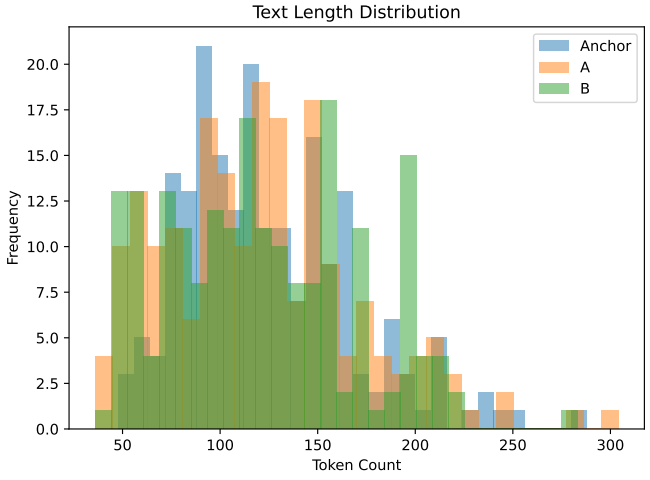

Fig. 1: Text Length Distribution. Most anchors, A, and B texts are 80–180 tokens, with similar distributions across all roles, reducing feature engineering needs. (Masab)

B. Boxplot of Text Lengths

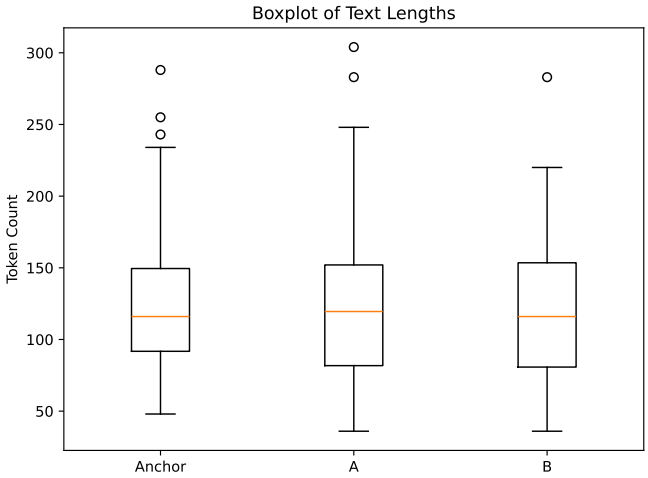

Fig. 2: Boxplot of Text Lengths. Medians at about 120 tokens for all types; outlier structure is also consistent. (Salar)

C. Vocabulary Size

Vocabulary observed: **14235 unique words**. (Ahmed)

D. Cosine Similarity Distribution

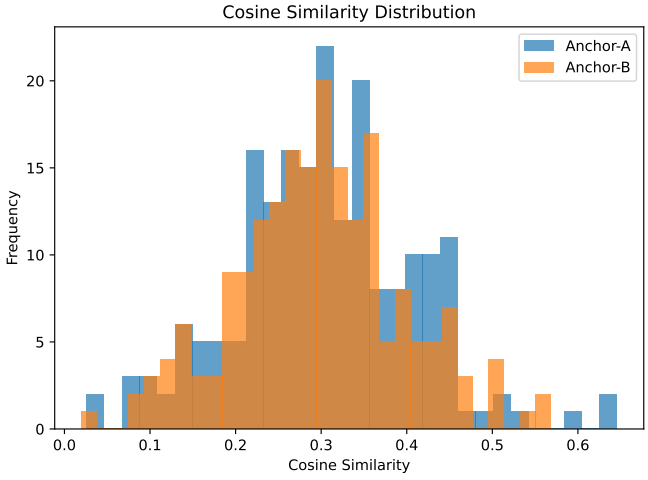

Fig. 3: Histograms of cosine similarity between anchor-candidate pairs. Overlap indicates semantic overlap and makes the task challenging for very close narratives. (Ahmed)

E. Difference in Similarity Scores

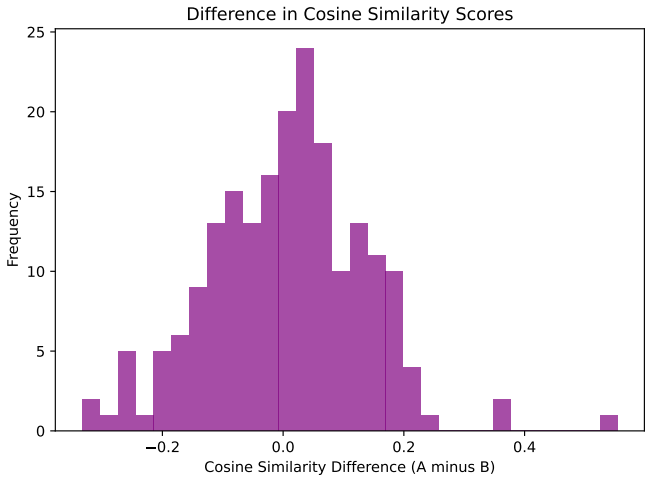

Fig. 4: Distribution of Anchor-A minus Anchor-B similarity scores. Extreme values highlight easy cases; scores near zero mark ambiguous triplets. (Masab)

F. Confusion Matrix

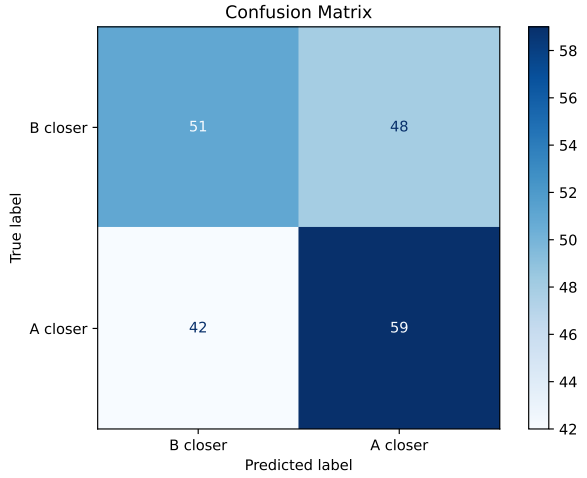

Fig. 5: Confusion matrix comparing cosine similarity rule to gold labels. The visible diagonal shows the cosine baseline works reasonably well. (Salar)

G. Class Balance Plot

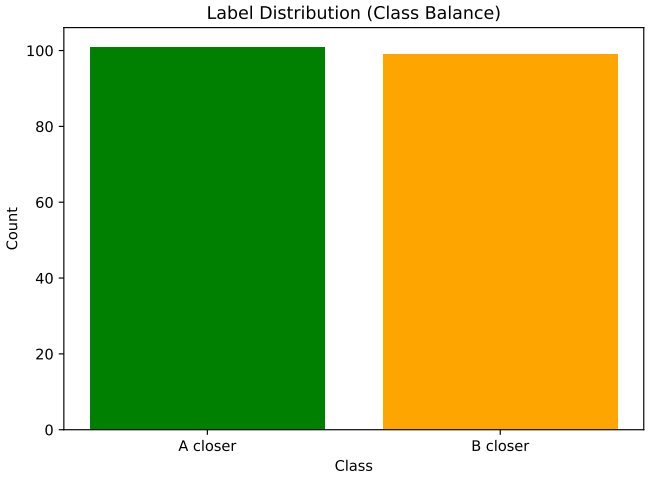

Fig. 6: Class balance for 'A closer' and 'B closer' is close to 1:1, minimizing risk of bias in downstream models. (Jahanzaib)

H. Sentence Length Distribution

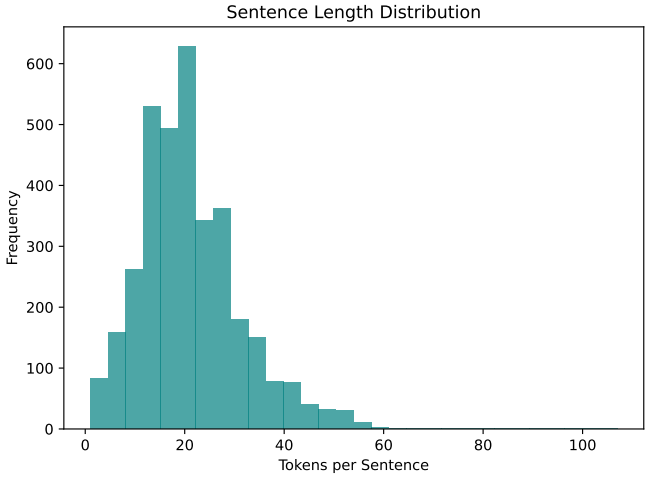

Fig. 7: Histogram of tokens per sentence (all anchor, A, B). Most sentences are short, though some long stories introduce right-tail outliers. (Masab)

I. t-SNE Embedding Visualization

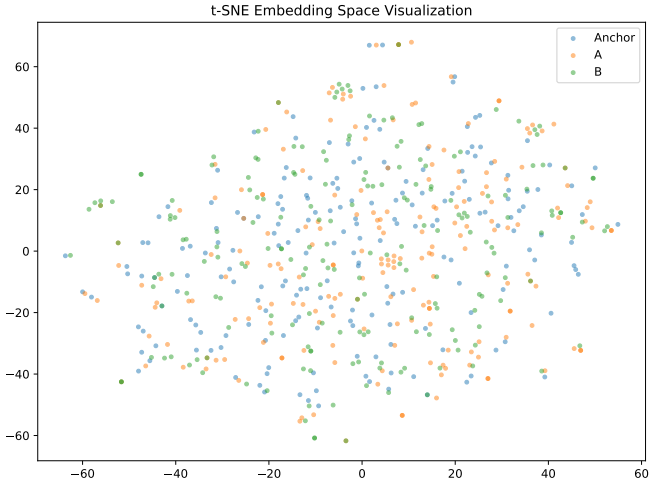

Fig. 8: t-SNE 2D projection of sentence embeddings for anchors, A, and B. Overlap suggests many narratives are not trivially separable using semantic vectors alone. (Ahmed)

IV. Task Distribution

TABLE I: CS 14B Group Task Assignment

Member	**Assigned Tasks**
Ahmed	Cosine similarity, t-SNE, vocab size
Masab	Histograms, difference, sentence lengths
Salar	Boxplots, confusion matrix
Jahanzaib	Class balance, report, tables

V. Challenges and Observations

No missing fields detected. Long stories (>250 tokens) appear as outliers but are rare. Class label balance and feature distribution support effective model development. Similarity-based baselines perform well for many pairs, but not all triplets.

VI. Conclusion

This EDA demonstrates that the narrative similarity Track A dataset is clean, consistent, and appropriately balanced for downstream modeling and experimentation.