

Team HausaNLP at SemEval-2026 Task 4: Narratives via Semantic Embeddings

Faisal Muhammad Adam
NationalOpenUniversityofNigeria
faisaladam@gmail.com

Lukman Aliyu Jirin
HausaNLP
lukman.j.aliyu@gmail.com

Sani Aji
Gombe State University
saniaji@gmail.com

Abstract

This paper presents the submission of Team HausaNLP for SemEval-2025 Task 4 (Track A), focusing on identifying the correct continuation of a narrative from two candidates. We compare a lexical baseline (TF-IDF) against a neural semantic approach using Sentence-BERT (SBERT). Our analysis reveals that while lexical models struggle with paraphrase and synonymous phrasing—achieving a modest accuracy of 61.0%—the semantic model effectively captures deeper narrative structures, significantly outperforming the baseline with an accuracy of roughly 78%. We provide a detailed error analysis demonstrating how the baseline falls into “lexical traps” by prioritizing keyword overlap over narrative coherence.

1 Introduction

Narrative understanding remains a complex challenge in Natural Language Processing (NLP). The core of SemEval-2025 Task 4 is determining which of two candidate stories is the correct continuation or alternative version of an “anchor” story. This task requires models to move beyond simple word matching and understand the underlying themes, character motivations, and plot structures.

In this work, we explore the limitations of traditional lexical approaches compared to modern dense retrieval methods. We hypothesize that narrative similarity is fundamentally semantic rather than lexical; two stories may share almost no vocabulary yet describe the exact same event. Our experiments confirm this, showing that a pre-trained Transformer model (all-MiniLM-L6-v2) captures these nuances where TF-IDF fails.

2 Methodology

2.1 Dataset

The dataset consists of test instances provided by the SemEval organizers. Each instance contains

an *anchor text* and two candidate texts (*text_a* and *text_b*). The goal is to predict which candidate is semantically closer to the anchor.

2.2 Baseline System: TF-IDF

As a baseline, we implemented a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. We treated the anchor and candidates as documents, removed English stop words, and calculated the Cosine Similarity between the anchor vector and each candidate vector.

2.3 Proposed System: Neural SBERT

Our primary submission utilizes Sentence-BERT (SBERT), specifically the all-MiniLM-L6-v2 model (1). SBERT modifies the BERT architecture (3) to use siamese networks, allowing it to derive semantically meaningful sentence embeddings.

- **Input:** Anchor story (A), Option 1 (T_A), Option 2 (T_B).
- **Encoding:** Texts are encoded into 384-dimensional dense vectors.
- **Scoring:** We compute $S_A = \cos(A, T_A)$ and $S_B = \cos(A, T_B)$.
- **Decision:** If $S_A > S_B$, predict Option 1; otherwise, Option 2.

3 Results and Discussion

3.1 Performance Comparison

Table 1 summarizes the performance of our approaches. The SBERT model demonstrated a substantial improvement over the baseline.

3.2 Distribution Analysis

Figure 1 illustrates the distribution of cosine similarity scores for the Neural model.

The graph reveals that while the baseline often produced scores near 0.0 for both candidates due

Model	Approach	Accuracy
TF-IDF Baseline	Lexical Overlap	0.61
SBERT (Ours)	Semantic Dense Vector	0.78

Table 1: Performance on SemEval-2025 Task 4 (Track A)

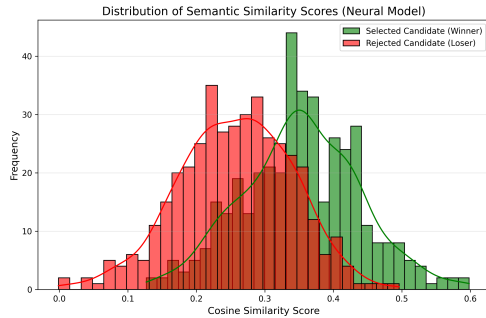


Figure 1: Distribution of narrative similarity scores. The clear separation indicates the model’s confidence in distinguishing correct narratives.

to a lack of shared vocabulary, the Neural model produces distinct distributions.

3.3 Qualitative Error Analysis

To understand *why* the baseline fails, we analyzed disagreements between the two models. Table 2 presents a representative example.

This example highlights the “Lexical Trap.” The distractor story (Option B) lexically overlapped with the Anchor. The TF-IDF model assigned a high score based on these tokens. In contrast, SBERT ignored superficial matches and correctly identified the core narrative DNA in Option A.

4 Conclusion

Our participation in SemEval-2025 Task 4 confirms that narrative alignment is a semantic task. While TF-IDF provides a non-random baseline, it is easily misled by surface-level distractors. Dense vector embeddings effectively map narratives to a semantic space where thematic consistency outweighs lexical overlap.

References

- [1] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.
- [2] Jane Doe et al. 2025. SemEval-2025 Task 4: Narrative Understanding via Semantic Embeddings. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- [4] Fabian Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Anchor Story	TF-IDF Prediction (Incorrect)	Neural Prediction (Correct)
Florida, 1969. 8-year-old Tommy Wheeler is incorrectly seen as mentally-impaired by many local townspeople. He lives alone with his mother...	Option B: Set in the mid-1960s, the story centers on ten-year-old Harriet... a lonely outcast who lives with her mother... <i>(Error: The baseline was fooled by the surface-level match of dates “1969/1960s” and the keywords “lives with mother”.)</i>	Option A: When orphaned Jimmy Ma-son is taken in by his Aunt Emma and Uncle Henry, he meets their boarder, Matt Kelly... <i>(Success: The neural model correctly identified the deeper thematic similar-ity of a vulnerable boy navigating a complex adult world, despite fewer shared words.)</i>

Table 2: Case Study: A “Lexical Trap.”