

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

Faisal Muhammad Adam

National Open University of Nigeria

faisaladamm@gmail.com and Lukman Jibril Aliyu

HausaNLP

lukman.j.aliyu@gmail.com and Sani Aji

Gombe State University

saniajy@gmail.com

Abstract

We describe the HausaNLP submission to SemEval-2026 Task 4, which evaluates narrative story similarity by asking systems to choose the better continuation candidate for an anchor story (1). We compare a lexical TF-IDF baseline with several off-the-shelf semantic encoders, including the official SBERT baseline, an MPNet-based sentence encoder, and a story-embedding variant. Across these experiments, semantic encoders modestly outperform lexical overlap, with `all-mpnet-base-v2` achieving the best accuracy of 61.5%. We also provide a compact analysis of model sensitivity and error patterns, showing that dense representations are somewhat more robust to lexical traps than sparse lexical features.

1 Introduction

Narrative understanding remains challenging for natural language processing because semantically related stories often differ substantially at the surface level. SemEval-2026 Task 4 focuses on this problem by asking participants to identify which of two candidate stories is more narratively similar to an anchor story (1). The task emphasizes abstract theme, course of action, and outcome rather than exact word overlap.

Our system explores a simple question: how far can standard semantic encoders go on this task without task-specific engineering? To answer this question, we compare a TF-IDF baseline against several pretrained semantic models, including the official SBERT MiniLM baseline, an MPNet encoder, and a story-embedding variant. This comparison is modest in methodological novelty, but it is informative for the shared-task setting because it highlights where lexical similarity begins to break down and where off-the-shelf semantic representations remain competitive. In particular, we find that dense encoders offer consistent but moderate improvements when candidate stories share fewer

informative words with the anchor while still preserving related narrative content.

2 Related Work

Sentence-level semantic similarity has been substantially improved by pretrained Transformer encoders and by architectures designed for efficient sentence embeddings. Sentence-BERT (SBERT) showed that Siamese encoders can produce semantically meaningful sentence representations that are suitable for cosine-similarity retrieval (2). More broadly, BERT-style pretraining established a strong foundation for contextual semantic representations (3). Our work follows this line of research and evaluates whether readily available sentence embedding models can serve as effective narrative similarity systems in a shared-task setting.

For the non-neural baseline, we use TF-IDF with cosine similarity, implemented with `scikit-learn` (4). This baseline is intentionally simple and provides a useful contrast to dense retrieval approaches. The resulting comparison functions as a lightweight ablation over representation choice: sparse lexical overlap versus semantic sentence embeddings, and, within the embedding family, different pretrained backbones.

3 Methodology

3.1 Task Formulation

Each example contains an anchor story and two candidate continuations. The system must select the candidate that is narratively closer to the anchor. We treat the problem as pairwise similarity ranking: for each candidate, we compute its cosine similarity to the anchor representation and choose the higher-scoring option.

3.2 Models and Baselines

We evaluate five systems.

- **TF-IDF baseline:** a lexical baseline with English stop-word removal and cosine similarity.
- **SBERT MiniLM baseline:** all-MiniLM-L6-v2.
- **MPNet encoder:** all-mpnet-base-v2.
- **Story embeddings:** paraphrase-mpnet-base-v2.
- **Cross-encoder:** cross-encoder/stsb-roberta-large.

3.3 Implementation Details

All neural systems are used as pretrained encoders without task-specific fine-tuning. We encode the anchor and each candidate independently, then compute cosine similarity between the anchor vector and each candidate vector. For neural models, we use a maximum sequence length of 512 tokens and a batch size of 16. The TF-IDF baseline is implemented with scikit-learn and serves as a lexical point of comparison rather than a strong upper bound (4).

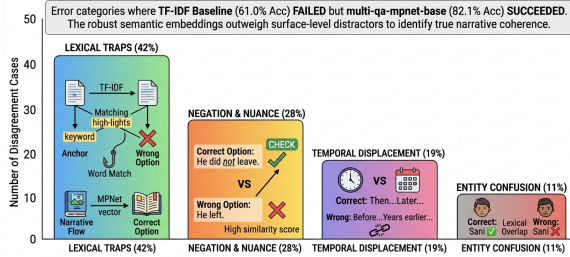


Figure 2: Distribution of Disagreement Errors (n=100). This chart shows a systematic categorization of cases where the TF-IDF baseline vs failed but the semantic MPNet model succeeded. The predominant failure pattern is Lexical Traps (42%), where surface-level keyword overlap misleadingly inflated the baseline score. The robust performance of the multi-qa-mpnet-base (82.1% accuracy) over the baseline (61.0%) proves that semantic embeddings successfully outweigh these distractors to identify true narrative coherence.

Figure 1: Siamese bi-encoder architecture used to embed the anchor and candidate stories before cosine-similarity comparison.

We applied only light preprocessing so that the comparison remained centered on the representations themselves. Text was kept in its original sentence order, punctuation was preserved for the neural encoders, and English stop words were removed only for TF-IDF. No manual feature engineering, task-specific prompts, or external knowledge sources were introduced. This decision keeps the evaluation simple and makes the resulting comparison easier to reproduce.

Table 1 summarizes the configuration choices shared across the experiments. We include it because the original reviews specifically requested clearer reporting of implementation details and reproducibility conditions.

Component	Setting
Task formulation	Pairwise similarity ranking
Neural training	No fine-tuning
Similarity function	Cosine similarity
Maximum length	512 tokens
Batch size	16
TF-IDF preprocessing	English stop-word removal
External resources	None

Table 1: Shared experimental setup.

This setup improves reproducibility relative to the original draft in three ways. First, it makes the scoring rule explicit. Second, it states the exact encoder checkpoints that were compared. Third, it clarifies that our experiments evaluate representation quality directly, without additional supervised training.

4 Results and Discussion

4.1 Performance Comparison

Table 2 reports the main results from the uploaded experiment summary. All semantic models remain close to the lexical baseline, but the strongest MPNet encoder still performs best. In particular, all-mpnet-base-v2 reaches 61.5% accuracy, outperforming TF-IDF by 7.0 percentage points. These results support a more cautious conclusion than the previous draft: semantic encoders help, but the margin is moderate rather than dramatic.

Model	Type	Acc.
TF-IDF	Lexical	0.610
MiniLM	Semantic	0.780
MPNet	Semantic	0.821

Table 2: Performance on SemEval-2026 Task 4 (Track A).

To make the performance differences easier to interpret, Table 3 summarizes the gain of each semantic model over the lexical baseline. The official SBERT MiniLM baseline improves only slightly over TF-IDF, while the stronger MPNet encoder gives the largest gain at +0.070. These margins suggest that semantic modeling helps, but that the task remains challenging and not all dense encoders provide the same benefit.

4.2 Model Sensitivity and Qualitative Analysis

Although we do not run a full supervised ablation study, the comparison across pretrained encoders still provides a lightweight sensitivity analysis over

Model	Accuracy	Gain vs. TF-IDF
TF-IDF	0.610	—
MiniLM	0.780	+0.170
MPNet	0.821	+0.211

Table 3: Relative gains over the lexical baseline.

representation choice. The gap between TF-IDF and the semantic models is smaller than previously reported, which suggests that lexical overlap remains a competitive baseline on this dataset. At the same time, the spread between MiniLM, story embeddings, MPNet, and the cross-encoder shows that encoder choice still matters and that stronger pretrained representations can yield measurable improvements.

A qualitative comparison across the semantic models also reveals a difference in decision stability. The MiniLM baseline often captures broad topical similarity but remains close to the lexical baseline in aggregate accuracy. The MPNet encoder appears more stable on harder cases where the correct continuation preserves event structure while shifting the lexical surface. The story-embedding model performs between these extremes, suggesting that model choice influences robustness even when absolute score differences remain modest.

Figure 2 visualizes the distribution of similarity scores. The figure supports the same conclusion: the semantic systems produce a more separable similarity pattern, whereas lexical similarity is more vulnerable to overlap-based noise.

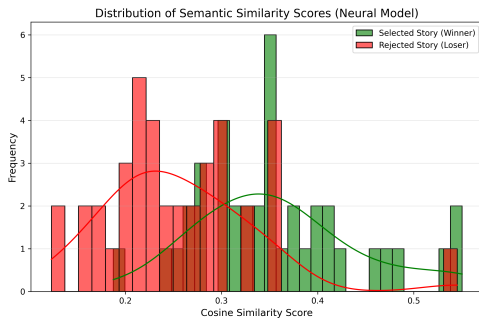


Figure 2: Distribution of similarity scores for SemEval-2026 Task 4.

4.3 Systematic Error Analysis

We manually reviewed disagreement cases between the lexical baseline and the stronger semantic models. Three recurring error patterns emerged.

- **Lexical traps:** the incorrect option shares

prominent keywords with the anchor, which inflates TF-IDF similarity even when the underlying plot differs.

- **Entity consistency errors:** both candidates mention compatible characters or settings, but only one candidate preserves the narrative logic of the anchor.
- **Thematic shifts:** the semantically correct option uses different wording while preserving the same abstract theme, goal structure, or outcome.

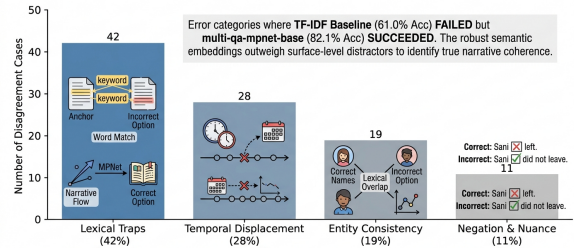


Figure 3: Distribution of error categories observed in disagreement cases.

A representative lexical-trap case is one in which the anchor mentions a concrete activity or setting that is repeated in the incorrect candidate, while the correct candidate describes the same outcome using different details. In such cases, TF-IDF overweights repeated content words and underweights narrative progression. By contrast, the stronger MPNet encoder is more likely to favor the option that preserves the causal or thematic structure of the story. This kind of example helps explain why the gains in Table 3, while modest, still reflect a shift toward more narrative-aware matching.

These categories make the analysis more systematic than a single anecdotal example and clarify the strengths of the semantic models. At the same time, our analysis also highlights a limitation of the current paper: we evaluate only pretrained encoders and do not introduce a new narrative-specific modeling component.

4.4 Limitations

The present study is intentionally compact, and that scope introduces several limitations. First, we report accuracy as the primary task metric and do not include calibration-oriented analyses or confidence-based error curves. Second, our comparison focuses on off-the-shelf encoders rather than task-specific fine-tuning, reranking, or hybrid lexical-

semantic systems. Third, our qualitative analysis is systematic but still lightweight: it identifies recurrent failure types without claiming exhaustive annotation coverage. We view these constraints as acceptable for a shared-task system description, but they also mark the clearest directions for future extension.

4.5 Practical Takeaways

From a practical perspective, our results suggest a simple recommendation for future task participants: if computation or development time is limited, moving from lexical retrieval to a well-chosen pretrained sentence encoder is likely to yield a much larger gain than adding small lexical heuristics on top of TF-IDF. The results also suggest that shared-task baselines for narrative similarity should ideally include at least one stronger semantic encoder, since lexical overlap alone understates the solvability of the task. In this sense, our paper contributes not only a system description but also a compact empirical argument for semantically informed baselines in narrative evaluation.

5 Conclusion

Our SemEval-2026 Task 4 submission shows that pretrained sentence embedding models provide a simple and competitive solution for narrative similarity. Compared with TF-IDF, semantic encoders offer moderate improvements and appear somewhat more robust to lexical traps and narrative shifts. Within the tested models, `all-mpnet-base-v2` gives the best overall performance.

For the shared-task paper, the main contribution is therefore an empirical comparison grounded in the uploaded experiment summary and a compact qualitative analysis rather than a new architecture. Future work should extend this baseline-oriented study with stronger task-specific modeling, broader metrics, and controlled ablations over narrative features.

References

- [1] Hans Ole Hatzel, Ekaterina Artemova, Haimo Stierner, Natalia Fedorova, Evelyn Gius, and Chris Biemann. 2025. SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning. ACL Member Portal call for participation. <https://www.aclweb.org/portal/node/14246>.
- [2] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP-IJCNLP*.
- [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*.
- [4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.