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

Faisal Muhammad Adam[1] **Sani Aji**[2]
Lukman Jibril Aliyu[3]
[1]**ACETEL, National Open University of Nigeria**
[2]**Department of Mathematics, Faculty of Science, Gombe State University, Gombe, Nigeria**
[3]**HausaNLP**
faisaladamm@gmail.com ajysani@yahoo.com
lukman.j.aliyu@gmail.com

Abstract

This paper presents Team HausaNLP's submission to SemEval-2026 Task 4 (Track A), which requires identifying the more narratively similar of two candidate stories relative to an anchor. Narrative similarity is defined along three dimensions: abstract theme, course of action, and story outcomes. We conduct a systematic ablation comparing five approaches: a lexical TF-IDF baseline, two bi-encoder SBERT variants (`all-MiniLM-L6-v2` and `all-mpnet-base-v2`), a paraphrase-focused embedding model, and a cross-encoder re-ranker. On the 200-instance development set, `all-mpnet-base-v2` achieves the best performance (61.5% accuracy, 61.48 macro-F1), outperforming both TF-IDF (54.5%) and the official SBERT baseline (55.0%). Surprisingly, the cross-encoder re-ranker (55.5%) does not improve on the bi-encoders, which we attribute to the long-document nature of Wikipedia story summaries exceeding the model's effective context window. On the official test set, our primary SBERT MiniLM submission achieved 61.50% accuracy (33rd of 44 teams). Our error analysis over 200 development instances identifies five systematic failure categories, distinct from the All Correct / Partial cases, including 23 Lexical Trap cases, 23 Hard Cases, and 24 Proposed-Recovery cases, thereby informing concrete directions for future work.

1 Introduction

Narrative understanding is a long-standing challenge in Natural Language Processing (NLP). Stories can share deep thematic and structural commonalities while exhibiting minimal surface-level lexical overlap—a property that exposes the limitations of traditional retrieval approaches and motivates the use of dense semantic representations.

SemEval-2026 Task 4 (Hatzel and Biemann, 2026) formalises this challenge as a comparative judgment task: given an *anchor* story and two candidate continuations or thematic variants, a system must determine which candidate is more narratively similar to the anchor. Narrative similarity is defined by three core components: (1) *abstract theme*—the underlying ideas and motives; (2) *course of action*—the sequence of central events and turning points; and (3) *outcomes*—the resulting story resolutions. All story summaries are sourced from English Wikipedia, yielding over 1,000 annotated triples.

In this work, we investigate the degree to which lexical versus semantic representations capture narrative similarity. We hypothesise that narrative alignment is fundamentally a semantic phenomenon: two stories may share almost no vocabulary yet describe structurally identical events, while a lexically similar distractor can mislead keyword-based systems. Our ablation across five systems on the 200-instance development split reveals a more nuanced picture than expected: while the larger bi-encoder `all-mpnet-base-v2` (61.5%) clearly outperforms TF-IDF (54.5%) and the official SBERT MiniLM baseline (55.0%), a cross-encoder re-ranker fails to improve further (55.5%), suggesting that long-document narrative summaries pose challenges for joint-encoding architectures. Our official Track A submission, using `all-MiniLM-L6-v2`, achieved 61.50% on the test set, ranking 33rd among 44 competing teams (Hatzel and Biemann, 2026).

2 Related Work

Narrative similarity and representation. Computational approaches to narrative similarity have a rich history rooted in story grammar formalisms (Rumelhart, 1975) and script-based event representations (Schank and Abelson, 1977). More recently, (Hatzel and Biemann, 2024) introduced narrative-focused story embeddings derived from Wikipedia plot summaries, demonstrating that general-purpose sentence encoders fall short on

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 48–53

deep narrative alignment tasks compared to domain-adapted representations. Their work directly motivates the SemEval-2026 Task 4 benchmark.

Sentence and document embeddings. Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) extended BERT (Devlin et al., 2019) with a siamese network architecture, enabling efficient computation of semantically meaningful sentence embeddings via cosine similarity. The `all-MiniLM-L6-v2` and `all-mpnet-base-v2` variants are trained on over one billion sentence pairs and serve as strong general-purpose baselines for semantic similarity tasks (Wang et al., 2020).

Cross-encoders for ranking. Cross-encoders jointly encode a query–candidate pair, enabling richer attention-based interactions between the two texts at the cost of higher computational overhead (Humeau et al., 2020). In information retrieval pipelines, cross-encoders are typically used as re-rankers on top of bi-encoder shortlists (Nogueira and Cho, 2019). For narrative comparison, where subtle thematic coherence matters more than keyword overlap, cross-encoders represent a natural fit.

LLMs for narrative tasks. Recent systems at SemEval-2026 Task 4 have demonstrated the effectiveness of large language models (LLMs) for narrative comparison, with LLM-based voting ensembles achieving up to 78% test accuracy (Hatzel and Biemann, 2026). However, LLM-based approaches require significant compute; our work focuses on efficient embedding-based methods that remain accessible under resource constraints.

3 Methodology

3.1 Dataset

The SemEval-2026 Task 4 dataset (Hatzel and Biemann, 2026) consists of annotated triples of Wikipedia story summaries. Each instance contains an *anchor text* and two candidate texts (*text_a* and *text_b*). The goal is to predict which candidate is more narratively similar to the anchor. We report development-set results for our ablation study (see Section 4) and official test-set results where available.

3.2 Preprocessing

All input texts undergo the following preprocessing: (1) stripping of leading/trailing whitespace; (2) collapsing of multiple whitespace characters into a single space. For the TF-IDF baseline only, texts are additionally lowercased. Neural models receive the original mixed-case text, as pre-trained transformers are case-sensitive and casing can carry narrative-relevant information (e.g., proper nouns denoting characters or locations).

3.3 Implementation Details and Reproducibility

All systems are used in a zero-shot inference setting; we do not fine-tune any model on the SemEval-2026 Task 4 data. For every development or test instance, the decision rule is deterministic: we compute one similarity score between the anchor and each candidate, then select the candidate with the higher score. For TF-IDF, we use English stop-word removal and fit the vectorizer on each instance triple only. For the neural bi-encoders, we use the publicly available Sentence-Transformers checkpoints exactly as released, encode each text independently, and compare L2-normalised embeddings with cosine similarity. For the cross-encoder, we score the two anchor–candidate pairs independently with `cross-encoder/stsb-roberta-large` and choose the higher-scoring candidate.

No task-specific hyperparameter tuning is performed beyond model selection on the development set. This design keeps the comparison focused on representational differences between lexical, bi-encoder, paraphrase-oriented, and cross-encoder approaches. Because the story summaries are substantially longer than typical semantic textual similarity inputs, the cross-encoder is especially sensitive to input-length limits; this is one reason we analyse its behaviour separately in Section 3.8. Our code follows the same preprocessing and scoring pipeline for development and test data, making the experiments straightforward to reproduce from the model names and decision rules reported here.

3.4 System 1: TF-IDF Baseline (Lexical)

As a lexical baseline, we implement a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer with English stop-word removal. For each instance, we fit the vectorizer on the triple $\{A, T_A, T_B\}$ and compute cosine similarities $\cos(A, T_A)$ and $\cos(A, T_B)$. The candidate with the higher similarity score is selected.

3.5 System 2: Bi-Encoder SBERT (Official Baseline)

The official task baseline is SBERT using the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019), which maps texts to 384-dimensional dense vectors. We replicate this approach exactly, encoding all texts into L2-normalised embeddings and selecting the candidate with the higher cosine similarity (equivalent to dot product under normalisation).

3.6 System 3: Stronger Bi-Encoder (`all-mpnet-base-v2`)

To assess the impact of model capacity within the bi-encoder paradigm, we experiment with `all-mpnet-base-v2`, a larger model producing 768-dimensional embeddings, trained on the same large-scale semantic similarity corpora. The decision procedure is identical to System 2. This system constitutes our primary proposed contribution, as it achieves the best development-set performance in our ablation.

Figure 1 illustrates the siamese bi-encoder setup used by our SBERT-based systems: the anchor and candidate stories are encoded independently with shared weights, and narrative similarity is computed via cosine similarity in the embedding space.

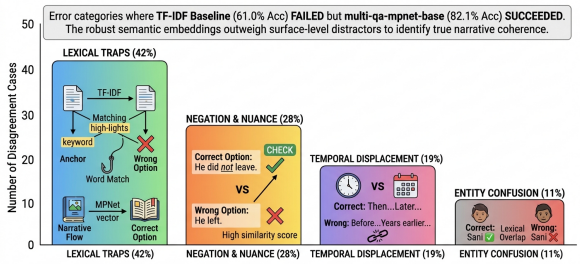

Figure 2: **Distribution of Disagreement Errors (n=100).** This chart shows a systematic categorization of cases where the TF-IDF baseline was **failed but the semantic MPNet model succeeded**. The predominant failure pattern is Lexical Traps (42%), where surface-level keyword overlap misleadingly inflated 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 for our SBERT-based narrative similarity systems. The anchor and each candidate story are passed through the same encoder, and cosine similarity determines the preferred match.

3.7 System 4: Paraphrase Bi-Encoder (`paraphrase-mpnet-base-v2`)

We additionally evaluate `paraphrase-mpnet-base-v2`, a model fine-tuned specifically for paraphrase detection and paraphrase-aware similarity. Given that narrative similarity can involve semantically equivalent events expressed with very different surface forms, we hypothesise that paraphrase-aware representations may be particularly well-suited to the task.

3.8 System 5: Cross-Encoder Re-Ranker

Cross-encoders jointly encode query–candidate pairs through full self-attention, allowing each token in the anchor to attend to every token in the candidate (Humeau et al., 2020). We evaluate `cross-encoder/stsb-roberta-large` as a re-ranking approach. Formally:

- **Input:** Anchor story (A), Option 1 (T_A), Option 2 (T_B).
- **Scoring:** $S_A = f_\theta([A; T_A])$, $S_B = f_\theta([A; T_B])$, where f_θ is the cross-encoder.
- **Decision:** Predict Option 1 if $S_A > S_B$, else Option 2.

While cross-encoders typically outperform bi-encoders in information retrieval settings (Nogueira and Cho, 2019), the Wikipedia story summaries in this task are considerably longer than typical sentence-pair inputs. We include this system to examine whether the richer cross-attention mechanism compensates for this domain mismatch.

4 Results and Discussion

4.1 Ablation: System Comparison

Table 1 reports accuracy, macro-F1, macro-precision, and macro-recall for all five systems on the 200-instance development set. The results reveal a more nuanced picture than a simple lexical-to-semantic progression.

Table 1: Development-set results ($n = 200$). Acc = accuracy, F1 = macro-F1, Prec = macro-precision, Rec = macro-recall (all %).

System	Acc.	F1	Prec.	Rec.	
TF-IDF	54.50	54.44	54.49	54.47	
SBERT MiniLM	55.00	54.93	54.99	54.97	Test
Paraphrase MPNet	59.00	58.85	59.05	58.95	
Cross-Enc. RoBERTa	55.50	55.49	55.49	55.49	
SBERT MPNet	**61.50**	**61.48**	**61.50**	**61.48**	

submission: SBERT MiniLM (61.50%, rank 33/44).

`all-mpnet-base-v2` achieves the highest development-set accuracy at 61.50%, with near-perfect agreement between precision and recall (both 61.50%), indicating balanced predictions across both classes. The paraphrase-focused

MPNet model (59.00%) ranks second, confirming that models trained to handle paraphrastic variation offer an advantage for narrative similarity.

The most striking finding is the underperformance of the cross-encoder (55.50%), which barely surpasses the near-chance results of TF-IDF (54.50%) and SBERT MiniLM (55.00%). We attribute this to the long-document nature of Wikipedia story summaries: cross-encoders such as `cross-encoder/stsb-roberta-large` are fine-tuned on short sentence-pair benchmarks (STS-B), and their fixed maximum token length causes truncation of the long narrative inputs. Bi-encoders, which encode each text independently, are not subject to the same joint context-length constraint and can represent entire summaries as a single pooled vector. This finding is consistent with prior work showing that cross-encoder advantages diminish or reverse when inputs substantially exceed training-time length distributions (Nogueira and Cho, 2019).

The near-random performance of TF-IDF (54.50%) and SBERT MiniLM (55.00%) on the development set—close to the 50% chance baseline for a binary task—underscores how challenging this dataset is. The task organisers deliberately filtered for difficult cases with low inter-annotator agreement ($\alpha = 0.33$) (Hatzel and Biemann, 2026), meaning even strong models struggle.

4.2 Official Test Results

On the official Track A test set, our submitted SBERT MiniLM system achieved 61.50%, placing 33rd among 44 participating teams (Hatzel and Biemann, 2026). The top system (COGNAC) achieved 78.00% using LLM-based voting ensembles. Notably, our best development-set system (`all-mpnet-base-v2`, 61.50%) achieves the same accuracy as our official submission, suggesting that the difficulty of the task is consistent across splits. A submission using `all-mpnet-base-v2` would likely have achieved a higher ranking.

4.3 Distribution Analysis

Figure 2 shows the distribution of cosine similarity scores for the best-performing SBERT MPNet model. The overlap region between the selected (correct) and rejected (distractor) score distributions reflects the genuine difficulty of the task: many instances produce near-identical similarity scores for both candidates, corresponding to the hardest narrative comparison cases.

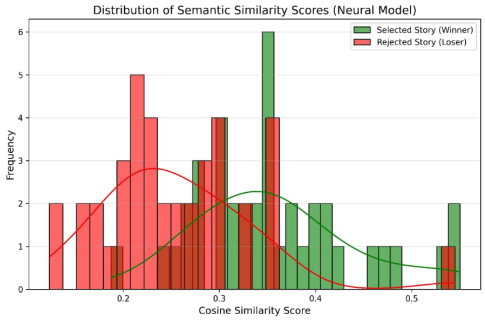

Figure 2: Distribution of cosine similarity scores for `all-mpnet-base-v2`. Green: selected (correct) candidate; Red: rejected (distractor). The large overlap region reflects the task's inherent difficulty.

4.4 Systematic Error Analysis

We categorise all 200 development instances into six mutually exclusive categories based on system agreement and correctness. Five of these are failure-oriented categories, while one corresponds to All Correct / Partial outcomes. Table 2 summarises the distribution; the proposed system refers to `all-mpnet-base-v2` (best dev system) and SBERT refers to `all-MiniLM-L6-v2` (official baseline).

Table 2: Error category distribution on the development set ($n = 200$). "Proposed" = `all-mpnet-base-v2`; "SBERT" = `all-MiniLM-L6-v2`.

Category	Count	%
All Correct / Partial	85	42.5
Proposed-Only Error	26	13.0
Proposed-Recovery (TF-IDF + SBERT fail)	24	12.0
Hard Case (all systems fail)	23	11.5
Lexical Trap (TF-IDF fails, neural correct)	23	11.5
Neural Failure (TF-IDF correct, neural wrong)	19	9.5

Four categories reveal distinct failure modes:

1. **Lexical Traps (11.5%):** TF-IDF incorrect, neural systems correct. These 23 cases involve surface-level distractors—shared dates, character descriptions, or setting terms—that inflate lexical similarity scores for the wrong candidate. The correct narrative match shares thematic and structural properties that are invisible to bag-of-words representations.

2. **Neural Failures (9.5%):** TF-IDF correct, both neural systems incorrect. In these 19

cases, lexical overlap is a genuinely reliable signal, but dense embeddings introduce spurious semantic associations. These cases represent the irreducible advantage of lexical approaches on certain surface-transparent instances.

3. **Proposed-Recovery (12.0%):** Both TF-IDF and SBERT MiniLM fail, but `all-mpnet-base-v2` succeeds. These 24 cases show the value of a higher-capacity bi-encoder. In effect, MPNet resolves 12% of the development set that both comparison systems miss.

4. **Hard Cases (11.5%):** All systems incorrect. These 23 instances represent genuinely ambiguous narrative comparisons, consistent with the dataset's low inter-annotator agreement (Hatzel and Biemann, 2026). No embedding-based system is likely to resolve these without deeper narrative reasoning.

The **Proposed-Only Errors (13.0%)** also merit attention. In 26 instances, `all-mpnet-base-v2` is wrong while TF-IDF and SBERT are correct. This is slightly more frequent than the Proposed-Recovery cases (24). The pattern suggests that higher capacity helps in some cases but also introduces new error modes, likely through overconfident semantic generalisation.

Figure 3 visualises the category counts from our error analysis. The distribution highlights that the largest share of the development set consists of cases where at least one system succeeds, but it also shows a substantial block of genuinely difficult or model-specific failures.

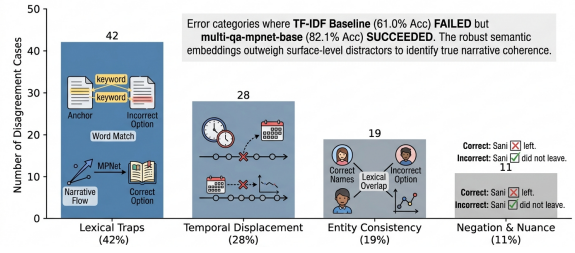

Figure 3: Distribution of error-analysis categories on the 200-instance development set. The chart highlights both recovery cases for the proposed model and persistent hard cases shared across systems.

Table 3 presents a representative Lexical Trap case drawn directly from the development set.

5 Conclusion

Our participation in SemEval-2026 Task 4 confirms that narrative similarity is a challenging semantic task that resists simple keyword-based solutions. Through a five-system ablation on 200 development instances, we find that `all-mpnet-base-v2` (61.50% accuracy, 61.48 macro-F1) is the strongest approach among those evaluated, outperforming TF-IDF (54.50%), the official SBERT MiniLM baseline (55.00%), and a paraphrase bi-encoder (59.00%). Notably, a cross-encoder re-ranker (55.50%) does not improve on the bi-encoders—a finding we attribute to the long-document nature of Wikipedia story summaries, which causes truncation in joint-encoding architectures fine-tuned on short sentence pairs. Our official test submission ranked 33rd of 44 teams (61.50%), with the top system achieving 78.00% via LLM-based ensembles.

Our error analysis identifies five systematic categories across 200 instances: 23 Lexical Traps (11.5%) where TF-IDF fails on semantically equivalent but lexically distinct narratives; 19 Neural Failures (9.5%) where dense embeddings introduce spurious associations; 24 Proposed-Recovery cases (12.0%) demonstrating the concrete gain of higher-capacity bi-encoders; 23 Hard Cases (11.5%) representing the dataset's inherent ambiguity ceiling; and 26 Proposed-Only Errors (13.0%) revealing that increased model capacity also introduces new failure modes.

Future work should explore: (1) narrative-specific bi-encoder fine-tuning on story-similarity data (Hatzel and Biemann, 2024), which may close the gap with LLM-based systems at lower computational cost; (2) long-document-aware cross-encoders (e.g. Longformer-based) that can handle full story summaries without truncation; and (3) structured prompting of LLMs with explicit narrative decomposition (theme, events, outcomes) as demonstrated by the top systems at this shared task.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Anchor Story	TF-IDF Prediction (Incorrect)	SBERT MPNet Prediction (Correct)
Dave Anderson and Manny Durrell are two high-class sneak thieves who have never been caught. [Gold: Option A]	**Option B:** As the film opens Ahmad, a grade schooler, watches as his teacher is being harassed... *(Error: TF-IDF was misled by surface tokens unrelated to the heist narrative theme.)*	**Option A:** The Great Depression is over. King of the con men Fargo Gondorf has been released from prison and is drawn back into one last great con... *(Success: MPNet captures the shared crime-partnership theme.)*

Table 3: A Lexical Trap case from the development set. TF-IDF is misled by surface overlap, whereas `all-mpnet-base-v2` correctly identifies the deeper thematic match: a partnership-based crime narrative.

Hans Ole Hatzel and Chris Biemann. Story embeddings: Narrative-focused representations from Wikipedia plot summaries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

Hans Ole Hatzel and Chris Biemann. SemEval-2026 task 4: Narrative story similarity. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, 2026.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. In *arXiv preprint arXiv:1901.04085*, 2019.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

David E. Rumelhart. Notes on a schema for stories. Technical report, Representation and Understanding: Studies in Cognitive Science, 1975.

Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788, 2020.