NLP-FSDM at SemEval-2026 Task 4: Narrative Similarity via Multiple Negatives Ranking and Instruction-Based Embeddings

Abdessamad Benlahbib[1], **Zouhir Essalmani**[1], **Achraf Boumhidi**[2], **Anass Fahfouh**[3], **Hamza Alami**[1]
[1] L3IA Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
[2]Department of Mathematics and Computer Sciences, National School of Applied Sciences Al Hoceima (ENSAH), UAE, Tetouan, Morocco
[3]Computer Science Department, Faculty of Sciences, UM5, Rabat
{abdessamad.benlahbib, zouhir.essalmani}@usmba.ac.ma,
{achraf.boumhidi, hamza.alami5}@usmba.ac.ma,
anassfahfouh@gmail.com

Abstract

The identification of narrative similarity is a complex NLP challenge that requires modeling deeper plot and thematic alignment rather than relying solely on lexical overlap. In this paper, we detail the participation of team NLP-FSDM in SemEval-2026 Task 4. Our approach utilizes the bge-large-en-v1.5 encoder. For Track A, we fine-tune it using Multiple Negatives Ranking Loss (MNRL), while for Track B we rely on the pretrained encoder to generate fixed narrative representations. We achieved an accuracy of 65.50% in Track A and 62.50% in Track B. This paper provides an extensive comparison of our results with competitive baselines and top-performing systems, analyzing the efficacy of dense encoders in low-resource narrative contexts.

1 Introduction

The computational modeling of narrative similarity has long struggled to move beyond simple keyword matching. While traditional Semantic Textual Similarity (STS) focuses on whether two sentences describe the same event, narrative similarity (NS) asks whether two stories share a similar "arc" or "trajectory" (Reimers and Gurevych, 2019; Hatzel and Biemann, 2024). The SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning (NSNRL) introduces a robust framework for testing these concepts using Wikipedia film plot summaries (Hatzel et al., 2026).

Narrative similarity, as defined by the organizers, focuses on abstract patterns of causality and progression while intentionally disregarding concrete details such as names, actors, objects, and specific settings (Hatzel et al., 2026). This "theory-agnostic" approach places a high demand on the model's ability to abstract away from the literal text to find the underlying story grammar (Rumelhart, 1975; Iyyer et al., 2016).

Our team, **NLP-FSDM**, entered both Track A (Comparative Similarity) and Track B (Representation Learning). For Track A, our strategy was built around the principle of contrastive learning (Chen et al., 2020a; Gao et al., 2021a), specifically leveraging the `bge-large-en-v1.5` (Xiao et al., 2023) model to learn a unified embedding space where stories with similar plot structures are clustered together. For Track B we relied on the pretrained `bge-large-en-v1.5` to generate fixed narrative representations. In the following sections, we detail our architectural choices, the training pipeline, and a comprehensive breakdown of our performance relative to the 48 participating teams. Our implementation code is available on GitHub.[1] The remainder of this paper is organized as follows: Section 2 provides an overview of related work in dense retrieval and narrative modeling. Section 3 formalizes the task definitions and specific challenges of narrative similarity. In Section 4, we detail our methodology, including data preparation, the Multiple Negatives Ranking Loss (MNRL) objective, and our hardware setup. Section 5 presents our official results and a comparative analysis with top-performing systems. Finally, Section 6 discusses the impact of subjectivity and limitations, followed by our concluding remarks in Section 7.

[1]Link to Source Code

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 612–616

2 Related Work

2.1 Sentence Embeddings and Dense Retrieval

Dense bi-encoder architectures have become a standard approach for semantic similarity (Benlahbib et al., 2024) and retrieval tasks. Sentence-BERT (Reimers and Gurevych, 2019) demonstrated that siamese BERT networks trained with contrastive objectives can produce high-quality sentence embeddings suitable for cosine similarity comparison. Subsequent work such as Dense Passage Retrieval (DPR) (Karpukhin et al., 2020a) further established bi-encoder models as efficient alternatives to cross-encoder architectures for large-scale retrieval.

2.2 Contrastive Learning for Text Representations

Contrastive learning has emerged as a powerful paradigm for representation learning across modalities. Methods such as SimCLR (Chen et al., 2020b) formalized the use of in-batch negatives, while SimCSE (Gao et al., 2021b) adapted contrastive objectives to sentence embeddings in NLP. Multiple Negatives Ranking Loss (MNRL), as implemented in SentenceTransformers, follows this principle by treating other examples within a batch as implicit negatives, enabling efficient learning without explicit hard-negative mining.

2.3 Narrative Modeling and Story Representation

Modeling narrative structure extends beyond surface semantic similarity. Prior work has explored plot units, event chains, and narrative arcs as structured representations of stories (Iyyer et al., 2016). The SemEval-2026 Task 4 dataset (Hatzel et al., 2026) evaluates whether modern embedding models can abstract away from lexical details to capture deeper thematic and structural similarity.

3 Task Definition and Challenges

SemEval-2026 Task 4 consists of two subtasks that evaluate the same core capability using different evaluation settings.

3.1 Track A: Comparative Similarity

In this track, systems are presented with a "triple": an anchor story $Anchor$, and two candidate stories A and B. The objective is to identify which candidate is more narratively similar to the anchor. This setup is conceptually simple but practically difficult because the dataset consists of summaries that have been filtered to ensure they are not "just a premise" but contain actual story components. Furthermore, the dataset was subjected to "rejection sampling," where only triples that caused disagreement between two commercial LLMs were kept, ensuring the benchmarks consist primarily of "hard cases".

3.2 Track B: Representation Learning

Track B requires systems to produce a fixed-length vector (embedding) for individual story instances. These representations are evaluated based on cosine distance: the distance between the anchor and the correct candidate must be smaller than the distance between the anchor and the distractor. This track is significantly more challenging as it forbids the use of cross-product information at inference time; the model must "understand" each story in isolation before any comparison occurs.

4 Methodology: The NLP-FSDM Approach

4.1 Data Filtering and Preparation

We utilized the provided training data, which consisted of 1,900 synthetic story triples. To ensure high data quality, we implemented a custom filter that removed any summaries missing essential fields or containing empty strings. This was crucial because the synthetic data occasionally contained incomplete or malformed entries, which could otherwise introduce noise into the training process. During inference, we applied instruction-based prefixes (“query:” and “passage:”) as recommended by the BGE framework:

- **Anchor Text:** Prefixed with *"query: "* to signal the intent for retrieval.
- **Candidates:** Prefixed with *"passage: "* to represent the search space.

4.2 Contrastive Fine-Tuning with MNRL

Our primary innovation was the use of **Multiple Negatives Ranking Loss (MNRL)** (Henderson et al., 2017; Reimers and Gurevych, 2019, 2020). Unlike standard Triplet Loss, which only considers one negative per anchor, MNRL treats every other story in a batch as a negative example. For a batch of size B, each anchor–positive pair is contrasted against the other positives in the batch, which serve as implicit negatives.

This objective is mathematically formulated as:

$$\mathcal{L} = -\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(a_i, p_i) \cdot \tau)}{\sum_{j=1}^{B} \exp(\text{sim}(a_i, p_j) \cdot \tau)} \quad (1)$$

where τ is a temperature hyperparameter. This loss function effectively forces the model to attend to the unique narrative features of the anchor that are shared with the positive candidate but absent in the "negatives" within the batch.

4.3 Hyperparameters and Training Protocol

We fine-tuned the `bge-large-en-v1.5` model for 5 epochs. We used a small batch size of 4 to accommodate the memory-intensive requirements of the 1024-dimensional BGE-Large encoder while maintaining a maximum sequence length of 512 tokens to capture the full narrative arc of the 4–8 sentence summaries. Our learning rate was set to $2e-5$ using the AdamW optimizer (Kingma and Ba, 2017) with a linear weight decay.

4.4 Model Usage Across Tracks

For Track A, we fine-tuned the `bge-large-en-v1.5` encoder using Multiple Negatives Ranking Loss as described above.

For Track B, however, we generated embeddings using the original pretrained `bge-large-en-v1.5` model without additional fine-tuning. This design choice allows us to assess the intrinsic narrative representation capacity of the base model independently from the contrastive fine-tuning used for comparative similarity.

4.5 Computing Environment and Hardware

All experiments, including the contrastive fine-tuning for Track A and the large-scale embedding extraction for Track B, were conducted using the Kaggle Kernels environment. We utilized a single NVIDIA Pascal P100 GPU with 16GB of VRAM. Due to the memory footprint of the `bge-large-en-v1.5` model (approximately 1.3GB in FP32), we optimized our training pipeline using gradient accumulation and small batch sizes to maintain stability within the P100's memory limits. The total training time for 5 epochs on the synthetic dataset was approximately 42 minutes, demonstrating the efficiency of the bi-encoder approach compared to more resource-intensive cross-encoder or LLM-based architectures.

5 Results and Performance Analysis

5.1 Official Results

As released by the organizers, Team **NLP-FSDM** achieved the following:

- **Track A Accuracy:** 65.50% (Rank 26)
- **Track B Accuracy:** 62.50% (Rank 16)

While our results fell short of the top-performing systems (which often relied on ensembling multiple commercial LLMs), we significantly outperformed the provided baselines. Specifically, in Track A, the Jaccard Similarity baseline (a lexical overlap measure) performed marginally above the random baseline (56.25% vs. 50.00%), while our system achieved a 15.5% gain over random guessing.

5.2 Comparative Analysis with Top Teams

A comparison with the top-scoring teams reveals a clear methodological divide:

- **COGNAC (Rank 1, Track A - 78.00%):** This system performed roughly on par with individual human annotators. The organizers noted that COGNAC, along with AI-Monitors and YNU-HPCC, relied on ensembling multiple commercial LLMs.
- **NLP-FSDM (Rank 26):** Our system represents a "dense retrieval" approach (Karpukhin et al., 2020b). Unlike the top teams, we did not use GPT-4o or other commercial APIs at inference time, making our system significantly more efficient and suitable for large-scale, offline corpus analysis.

In Track B, the drop in overall system performance was universal. The best system (COGNAC) achieved 72.00%, a 10-point drop from the Track A lead. This confirms the "bi-encoder bottleneck" where representing narrative nuance in a single vector is inherently harder than a direct side-by-side comparison. NLP-FSDM's Rank 16 in Track B reflects the representational capacity of the pretrained BGE-large model in modeling narrative similarity without task-specific fine-tuning.

6 Discussion

6.1 The Impact of Subjectivity

The low inter-annotator agreement (Krippendorff's alpha of 0.33) highlights the subjective nature of

the task. The organizers noted that human agreement was close to random chance for "Abstract Theme" and "Course of Action" (0.05 and 0.07 respectively). Despite this subjectivity, our system performs substantially above the random baseline, suggesting that dense encoders can capture latent narrative regularities beyond surface lexical cues.

6.2 Strengths and Limitations

Our use of MNRL allowed the model to focus on "hard negatives" within the training batches. However, our reliance on a single encoder limited our ability to capture the "Aspects" of similarity (Action, Outcome, Theme) that more complex ensembled systems could weigh. Future work should involve a hybrid approach where a dense encoder like BGE is augmented with structural features or plot units as defined in narrative theory.

7 Conclusion

The NLP-FSDM participation in SemEval 2026 Task 4 demonstrates that instruction-tuned dense encoders are a viable, efficient alternative to LLM-heavy ensembles for narrative similarity. While our contrastively fine-tuned model improved performance in Track A, Track B results demonstrate that the pretrained BGE-large encoder already captures substantial narrative structure even without additional supervision.

References

Abdessamad Benlahbib, Anass Fahfouh, Hamza Alami, and Achraf Boumhidi. 2024. NLP-LISAC at SemEval-2024 task 1: Transformer-based approaches for determining semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 213–217, Mexico City, Mexico. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021a. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2025 task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA.

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings — narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

David E. Rumelhart. 1975. Notes on a schema for stories. In DANIEL G. BOBROW and ALLAN COLLINS, editors, *Representation and Understanding*, pages 211–236. Morgan Kaufmann, San Diego.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.