ClaimCatchers at SemEval-2025 Task 7: Sentence Transformers for Claim Retrieval

Rrubaa Panchendrarajan[*]
Queen Mary University of London
r.panchendrarajan@qmul.ac.uk

Rafael Martins Frade[*]
University of Santiago de Compostela
Newtral
rafael.martins@newtral.es

Arkaitz Zubiaga
Queen Mary University of London
a.zubiaga@qmul.ac.uk

Abstract

Retrieving previously fact-checked claims from verified databases has become a crucial area of research in automated fact-checking, given the impracticality of manual verification of massive online content. To address this challenge, SemEval 2025 Task 7 focuses on multilingual previously fact-checked claim retrieval. This paper presents the experiments conducted for this task, evaluating the effectiveness of various sentence transformer models—ranging from 22M to 9B parameters—in conjunction with retrieval strategies such as nearest neighbor search and reranking techniques. Further, we explore the impact of learning context-specific text representation via finetuning these models. Our results demonstrate that smaller and medium-sized models, when optimized with effective finetuning and reranking, can achieve retrieval accuracy comparable to larger models, highlighting their potential for scalable and efficient misinformation detection.

1 Introduction

Given the vast volume of online content, manually fact-checking every claim is impractical and time-consuming. Moreover, many claims are recurrent, which do not need to be verified repeatedly (Shaar et al., 2020a). To address these challenges, retrieving previously fact-checked claims from a database can greatly support an automated fact-checking pipelines (Panchendrarajan and Zubiaga, 2024).

Research on retrieving previously fact-checked claims (claim retrieval) has evolved from traditional methods, such as BM25 (Robertson et al., 2009), to more advanced approaches leveraging sentence transformers (Reimers and Gurevych, 2019) and language models (Shaar et al., 2020a; Choi and Ferrara, 2024; Pisarevskaya and Zubiaga, 2025). However, given the dearth of research tackling the task from a multilingual angle, SemEval 2025 Task 7 (Peng et al., 2025) focuses on multilingual claim retrieval. Given a social media post, the task aims to retrieve a matching claim from a database of fact-checked claims. The task features two challenges: monolingual (posts and claims in the same language) and cross-lingual (posts and claims in different languages) claim retrieval.

We present our experiments for both the monolingual and cross-lingual tasks. We investigate the effectiveness of various sentence transformer models, ranging from 22M to 9B parameters, in combination with retrieval strategies such as nearest neighbors and reranking techniques. Additionally, we finetune these models to assess the impact of learning context-specific representation on claim retrieval tasks. Our findings reveal that smaller and medium-size models can achieve performance comparable to larger models when optimized with effective finetuning and reranking strategies. This suggests that these computationally effective models can still be adapted for accurate retrieval, making them a practical choice for real-world automated fact-checking systems.

2 Background

One of the first studies on claim retrieval was Shaar et al. (2020a), who introduced a pipeline using a fast lexical search to select fact-checked claims, followed by language models for reranking.

The claim retrieval task also featured in the 2020-2022 editions of the CLEF-CheckThat competition (Shaar et al., 2020b, 2021; Nakov et al., 2022). In 2020, the best-performing team, Buster.ai, used data augmentation and additional training datasets (Bouziane et al., 2020). In 2021, the best result in the English task was obtained by team Aschern (Chernyavskiy et al., 2021), which fine-tuned SBERT (Reimers and Gurevych, 2019) and then used LambdaMART to rank the top 20 matches. Finally, in 2022, team RIET ranked first (Shlisel-

[*] These authors contributed equally.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 455–462
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

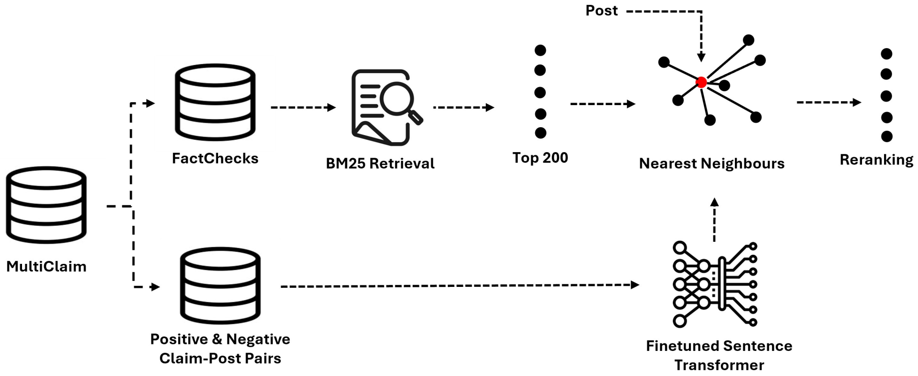


Figure 1: System Overview

berg and Dori-Hacohen, 2022) using Sentence-T5 for candidate selection, instead of BM25. For final reranking, they utilized GPT-Neo, an autoregressive language model (Black et al., 2021). Using LLMs to rerank selected candidates or match claim/post pairs has also been tried (Qin et al., 2023; Choi and Ferrara, 2024).

The task of retrieving previously fact-checked claims has been carried out primarily using datasets in English. Some of the exceptions were the 2021 edition of CheckThat (Shaar et al., 2021) which contained a claim retrieval subtask with Arabic tweet/claim pairs and a dataset presented by Kazemi et al. (2021) with WhatsApp and public group messages in English, Hindi, Bengali, Malayalam and Tamil. The currently most comprehensive dataset with multilingual previously fact-checked claims is MultiClaim (Pikuliak et al., 2023), with 28k posts in 27 languages from social media and 206k fact-checks in 39 languages.

3 System overview

This section discusses our claim retrieval approaches, nearest neighbors search and reranking. We also enhance retrieval by finetuning sentence transformers on training data. Figure 1 illustrates our method.

The task organizers provided social media posts with original text, English translations, and OCR-extracted text (if available), while fact-checks included original and translated text. Given a social media post, we concatenated its translated text and OCR-translated text to form the query for retrieving fact-check translations.

3.1 Nearest Neighbors

The dominant strategy in modern information retrieval aims to combine language model representations with efficient search (Zhao et al., 2024). This approach consists of using language models to generate vector representations for the documents and queries, and then performing the vector search with approximate nearest neighbors (ANN) algorithms. The most widely used ANN algorithm is Hierarchical Navigable Small Words (HNSW) (Malkov and Yashunin, 2018). HNSW works based on a layered graph structure that balances connectedness and shortest path length.

That was the first approach taken for the Semeval task. We generated vector representations for both claims and posts using sentence transformer models of varying sizes, and then HNSW to retrieve the claims for each post.

3.2 Reranking

As seen in the background section, early research on claim retrieval primarily relied on the lexical search to retrieve a small subset of claims, followed by reranking using embeddings generated by a sentence transformer model. This method has the advantage of limiting the use of computationally expensive resources to only a small portion of the data. A potential drawback is that the final result will depend on the selected samples.

We used an optimized version of BM25 (Lù, 2024) to select the top 200 candidates in the reranking approach. Since BM25, a bag-of-words method, is sensitive to noise, we preprocessed text by lowercasing, removing non-alphabetic characters, and stemming. In contrast, for reranking, we generated vector representations from the original text, as language models handle uncleaned text effectively. Finally, the top 200 candidates were ranked based on L2 similarity.

Group	Model	Parameters	Embedding Size	Context Length	Language
Small	sentence-transformers/all-MiniLM-L6-v2	22M	384	512	English
	sentence-transformers/all-mpnet-base-v2	278M	768	514	English
	avsolatorio/GIST-large-Embedding-v0	335M	1024	514	English
	sentence-transformers/all-roberta-large-v1	355M	1024	514	English
Medium	NovaSearch/stella_en_400M_v5	435M	4096	8192	English
	HIT-TMG/KaLM-embedding-multilingual-v1.5	494M	896	13K	Multilingual
	intfloat/multilingual-e5-large	560M	1024	514	Multilingual
Large	Alibaba-NLP/gte-Qwen2-1.5B-instruct	1B	8960	32K	Multilingual
	Alignment-Lab-AI/e5-mistral-7b-instruct	7B	4096	32K	Multilingual
	BAAI/bge-multilingual-gemma2	9B	3584	8192	Multilingual

Table 1: Sentence Transformer Models Used

3.3 Fine-tuned models

Finetuning language models on domain-specific data has been widely recognized as effective in the literature (Choi and Ferrara, 2024), as it enables models to learn context-specific sentence representations. To enhance retrieval performance, we finetune sentence transformer models using the Train partition. Sentence transformers (Reimers and Gurevych, 2019) support datasets representing various notions of textual similarity. For finetuning, we adopt the *pair-class* approach, where each training instance consists of a premise, a hypothesis, and a label indicating their similarity.

We used the 25K claim-post pairs from the Train partition as the positive samples. An equal amount of negative samples were generated by randomly selecting claim-post pairs that are not annotated as true pairs in the Train partition. While this method may occasionally introduce false negatives, the likelihood is minimal given the ample number of claims and posts in the dataset. This process resulted in a training set of 51K samples. Sentence transformer models were finetuned using this training set and then leveraged to obtain the sentence representation for the Nearest neighbor and reranking approaches discussed earlier.

4 Experimental setup

4.1 Sentence Transformer Models

We experiment with various sentence transformer models for both monolingual and cross-lingual retrieval tasks. We categorize them into the following three groups based on the number of parameters.

- Small: Fewer than 400 M parameters
- Medium: Between 400M and 1B parameters
- Large: More than 1B parameters

Table 1 lists the models used in the experiments and their characteristics. They range in size from 22M to 9B parameters, with embedding dimensions between 384 and 8960, and context lengths from 512 to 32K. We experiment with monolingual and multilingual models. The largest model in our study, *BAAI/bge-multilingual-gemma2* (Chen et al., 2024), is based on *google/gemma-2-9b* (Team, 2024) and contains 9B parameters. This model was trained on a diverse dataset spanning multiple languages and tasks, including retrieval. All the models used in the experiments are available in Huggingface via the Sentence Transformer library (Reimers and Gurevych, 2019).

5 Results

5.1 Environment Setting

Although the medium and large models support a higher context length, we restrict the maximum text length for obtaining sentence embeddings using sentence transformers to 512. This limitation is primarily due to memory constraints, as higher context lengths often result in out-of-memory errors. However, we were unable to fine-tune large models, as they require additional memory optimization techniques, such as Low-Rank Adaptation (LoRA), to enable efficient fine-tuning with limited GPU memory. Consequently, we fine-tuned only the small and medium-sized models. All experiments were performed using a single GPU (either Volta V100 or Ampere A100) with 8 CPU cores and 11 GB of memory per core for training and testing all models. The other hyperparameters used across retrieval and finetuning processes are discussed in Appendix A.1.

We report Success@10, which measures the fraction of queries where the corresponding fact-check was retrieved within the top 10 results in the train and development data released by the task organizers. For the monolingual task, we compute the language-wise average.

		Monolingual		Cross-lingual	
Approach	Model	Train	Dev	Train	Dev
Nearest Neighbours	sentence-transformers/all-MiniLM-L6-v2	0.779	0.754	0.57	0.549
	sentence-transformers/all-mpnet-base-v2	0.754	0.746	0.554	0.569
	avsolatorio/GIST-large-Embedding-v0	0.849	0.828	0.691	0.665
	sentence-transformers/all-roberta-large-v1	0.759	0.753	0.555	0.544
	NovaSearch/stella_en_400M_v5	0.84	0.822	0.693	0.678
	HIT-TMG/KaLM-embedding-multilingual-v1.5	0.82	0.791	0.637	0.624
	intfloat/multilingual-e5-large	0.833	0.83	0.66	0.644
	Alibaba-NLP/gte-Qwen2-1.5B-instruct	0.814	0.802	0.64	0.651
	Alignment-Lab-AI/e5-mistral-7b-instruct	0.788	0.776	0.634	0.638
	BAAI/bge-multilingual-gemma2	**0.879**	**0.881**	**0.74**	**0.716**
Reranking	sentence-transformers/all-MiniLM-L6-v2	0.801	0.786	0.596	0.587
	sentence-transformers/all-mpnet-base-v2	0.787	0.773	0.591	0.591
	avsolatorio/GIST-large-Embedding-v0	0.842	0.828	0.644	0.624
	sentence-transformers/all-roberta-large-v1	0.795	0.787	0.597	0.598
	NovaSearch/stella_en_400M_v5	0.839	0.825	0.642	0.636
	HIT-TMG/KaLM-embedding-multilingual-v1.5	0.827	0.801	0.63	0.607
	intfloat/multilingual-e5-large	0.835	0.833	0.639	0.629
	Alibaba-NLP/gte-Qwen2-1.5B-instruct	0.827	0.81	0.634	0.622
	Alignment-Lab-AI/e5-mistral-7b-instruct	0.812	0.799	0.622	0.616
	BAAI/bge-multilingual-gemma2	0.861	0.854	0.655	0.642

Table 2: Performance of the Pretrained Models on Nearest Neighbors and Reranking Approaches

		Monolingual		Cross-lingual	
Approach	Model	Train	Dev	Train	Dev
Nearest Neighbours	sentence-transformers/all-MiniLM-L6-v2	0.763	0.725	0.558	0.542
	sentence-transformers/all-mpnet-base-v2	0.822	0.783	0.641	0.607
	avsolatorio/GIST-large-Embedding-v0	0.832	0.769	0.665	0.618
	sentence-transformers/all-roberta-large-v1	0.836	0.784	0.669	0.598
	NovaSearch/stella_en_400M_v5	**0.88**	0.818	**0.734**	**0.653**
	HIT-TMG/KaLM-embedding-multilingual-v1.5	0.86	0.777	0.683	0.595
	intfloat/multilingual-e5-large	0.853	0.802	0.688	0.625
Reranking	sentence-transformers/all-MiniLM-L6-v2	0.79	0.764	0.582	0.549
	sentence-transformers/all-mpnet-base-v2	0.832	0.802	0.626	0.591
	avsolatorio/GIST-large-Embedding-v0	0.837	0.797	0.638	0.591
	sentence-transformers/all-roberta-large-v1	0.844	0.816	0.64	0.595
	NovaSearch/stella_en_400M_v5	0.859	**0.833**	0.655	0.615
	HIT-TMG/KaLM-embedding-multilingual-v1.5	0.852	0.808	0.64	0.584
	intfloat/multilingual-e5-large	0.846	0.82	0.643	0.6

Table 3: Performance of the Finetuned Models on Nearest Neighbors and Reranking Approaches

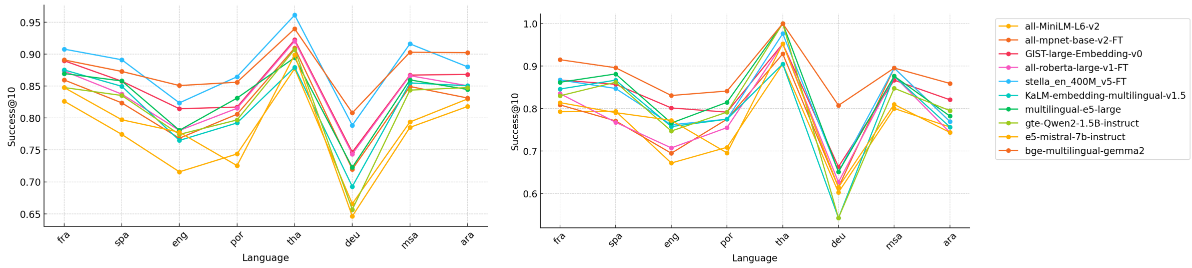
5.2 Performance of Pretrained Models

Table 2 presents the Success@10 scores of various models. Notably, the largest model, *bge-multilingual-gemma2* (Chen et al., 2024), outperforms all other models in both monolingual and cross-lingual tasks. However, most of the medium-sized or large models do not benefit from the reranking approach. This suggests that these models are already powerful enough to retrieve relevant results within the top 200, making the combination with BM25 in a two-step ranking approach ineffective. In contrast, among the smaller models, reranking consistently improves performance—except for *GIST-large-Embedding-v0* (Solatorio, 2024)—with the highest improvement of 5.4% observed with the *all-roberta-large-v1 model* (Liu et al., 2019) model in the cross-lingual task.

Interestingly, *GIST-large-Embedding-v0* (Solatorio, 2024) from the small group and *stella_en_400M_v5* (Zhang et al., 2024) from the medium group outperform most of the medium-sized and large models. This highlights that, despite having fewer parameters, these models are highly effective in representing text as embeddings. We submitted the predictions generated by the pretrained *bge-multilingual-gemma2* (Chen et al., 2024) model using the nearest neighbor approach to the task leaderboard, as this method demonstrated the best performance. Our submission achieved 18th place in the monolingual task and 15th place in the crosslingual task.

5.3 Performance of Finetuned Models

In this section, we discuss the impact of the finetuning process on retrieval approaches. As men-



(a) Train

(b) Dev

Figure 2: Language-wise Performance of the Models

tioned earlier, we fine-tuned only the small and medium-sized sentence transformers. Table 3 presents the performance of the fine-tuned models in both nearest neighbor and reranking approaches. Among the models compared, *all-mpnet-base-v2*, *all-roberta-large-v1* and *stella_en_400M_v5* show significant performance gains in *Train* and *Dev* partitions. Notably, *all-roberta-large-v1* (Liu et al., 2019) achieves the highest improvement, with an 11.4% increase in the cross-lingual task. Furthermore, the fine-tuned *stella_en_400M_v5* model attains performance competitive with the largest model, *bge-multilingual-gemma2* (Chen et al., 2024), in both tasks on the *Train* partition. This demonstrates that with further focus on generalizing these models, it is possible to achieve powerful sentence transformer representations comparable to those of larger models. Similar to the behavior of pretrained models, the reranking approach benefits only the smaller models. However, this trend is observed exclusively in the monolingual task, while reranking with fine-tuned models tends to decrease performance in cross-lingual task.

5.4 Language-wise Performance

We analyze the language-wise performance of the pretrained and finetuned models in the nearest neighbor approach. We choose either a pretrained or finetuned model depending on its performance in the Train partition. Figure 2 presents the Success@10 scores across languages, with finetuned models denoted by the postfix *FT*.

All models exhibit strong performance in the *Thai* language, followed by *Malay, French, Spanish* and *Arabic*. Conversely, *German* demonstrates the weakest retrieval performance. Despite being a high-resource language, *English* achieves only moderate performance. These trends, however, are influenced by factors such as imbalanced language samples, translation errors, and data annotation issues (discussed further in Appendix A.2). Among the models evaluated, *bge-multilingual-gemma2* consistently achieves the highest performance across languages in the Dev partition. While the medium-sized model *stella_en_400M_v5* performs best in Train, its performance declines in Dev, suggesting limited generalizability.

6 Conclusion

This paper presents the experiments conducted by the ClaimCatchers team for SemEval task 7, focusing on multilingual fact-checked claim retrieval. We investigate the performance of various sentence transformer models, categorized into three groups—small, medium, and large—based on the number of parameters. These models are evaluated using a range of approaches, including nearest neighbor, reranking, and fine-tuning.

Our experiment results show that the largest pretrained model, *BAAI/bge-multilingual-gemma2* yields the best performance when applied with nearest neighbor approach in both monolingual and cross-lingual tasks. Furthermore, the pretrained medium-sized and large models are powerful enough to retrieve more accurate results than when combined with reranking strategies like BM25. Interestingly, some small and medium-sized pretrained models outperform the larger models, indicating that the number of parameters is not the sole factor in achieving rich sentence representations. Additionally, the finetuned medium-sized model, *NovaSearch/stella_en_400M_v5*, is competitive with the largest model in *Train* partition, indicating that smaller or medium-sized models, with focused fine-tuning, can achieve performance comparable to larger models. These findings highlight the potential of smaller and fine-tuned models to achieve competitive performance, emphasizing the importance of selecting the appropriate ranking and fine-tuning strategies.

Acknowledgments

Rafael Martins Frade and Rrubaa Panchendrarajan are funded by the European Union and UK Research and Innovation under Grant No. 101073351 as part of Marie Skłodowska-Curie Actions (MSCA Hybrid Intelligence to monitor, promote, and analyze transformations in good democracy practices). We acknowledge Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT, for enabling our experiments (King et al., 2017).

References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with meshtensorflow.

Mostafa Bouziane, Hugo Perrin, Aurélien Cluzeau, Julien Mardas, and Amine Sadeq. 2020. Team buster. ai at checkthat! 2020 insights and recommendations to improve fact-checking. In *CLEF (Working Notes)*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2021. Aschern at checkthat! 2021: lambda-calculus of fact-checked claims. *Faggioli et al.[12]*.

Eun Cheol Choi and Emilio Ferrara. 2024. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1441–1449.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A Hale. 2021. Claim matching beyond english to scale global fact-checking. *arXiv preprint arXiv:2106.00853*.

Thomas King, Simon Butcher, and Lukasz Zalewski. 2017. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *arXiv preprint arXiv:2407.03618*.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podroužek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. Semeval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smolen, Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. *arXiv preprint arXiv:2305.07991*.

Dina Pisarevskaya and Arkaitz Zubiaga. 2025. Zero-shot and few-shot learning with instruction-following llms for claim matching in automated fact-checking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9721–9736.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

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*. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, Giovanni Da San Martino, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the clef-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In *CLEF (Working Notes)*, pages 393–405.

Shaden Shaar, Giovanni Da San Martino, Nikolay Babulkov, and Preslav Nakov. 2020a. That is a known lie: Detecting previously fact-checked claims. *arXiv preprint arXiv:2005.06058*.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeno, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, et al. 2020b. Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media. *CLEF (Working Notes)*, 2696.

SD-H Michael Shliselberg and Shiri Dori-Hacohen. 2022. Riet lab at checkthat! 2022: improving decoder based re-ranking for claim matching. *Working Notes of CLEF*, pages 05–08.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

Gemma Team. 2024. Gemma.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

A Appendix

A.1 Hyperparameters

Table 4 lists the hyperparameters used across different approaches.

	Hyperparameter	Value
Nearest Neighbour	Distance	l2
	No. bidirectional links per node	48
	No. neighbors considered during graph construction	200
	No. neighbors considered during search	200
BM25	Document length normalization factor	1
Sentence Transformers	Maximum sequence length	512
Finetuning	Batch size	16
	Learning rate	2E-05
	Warmup ratio	0.1
	No. epochs	1

Table 4: Hyperparamets

A.2 Data Labeling Issues

Analysing the data, we observed cases where a fact-check that appears among the top 10 candidates for a post could be considered correct, but it's not mapped as a valid pair. It is natural to expect that there are similar posts and fact-checks in which a small difference is enough to make the pair invalid, but we found cases where potentially valid pairs are not labeled as such, as shown in Table 5. This could explain part of the limited performance of smaller language models: they can identify very close matches, but these matches are not recognized as valid results.

Figure 3 presents the distribution of the cosine distance between valid pairs and posts computed using sentence-transformers/all-MiniLM-L6-v2. With a mean of 0.34, the distribution indicates that smaller cosine distances increase the probability of a retrieved fact-checked being valid. However, among all retrieved results in the top 10 with a cosine distance lower than 0.2, only 40% are mapped as valid pairs. While the remaining 60% likely include incorrect post/fact-check pairs, some of them could potentially be considered correct.

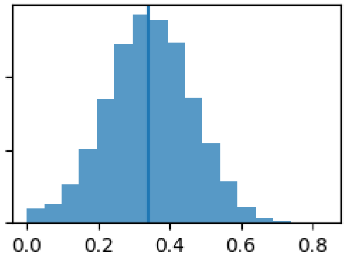


Figure 3: Distribution of cosine distance of valid post/fact-check pairs. With mean value of 0.34

This supports the organizers' decision to use Success@10 as the evaluation metric, as it mitigates the impact of misclassified pairs that should have been mapped and are likely to appear among the top results.

Post ID	**Post Text**	**Fact-check ID**	**Fact-check Text**	**Cosine Distance**
6228	In Brazil, a corrupt city councilor, tied to a pole by the population...	27741	In Brazil, a corrupt city councilor was tied to a pole by his fellow citizens.	0.08
13009	VLADIMIR PUTIN'S HOUSE IN SOCHI...	80938	Putin's house in Sochi	0.07
69	#URGENT More Madrid intends close the Pandemic Hospital Isabel Zendal if she wins the elections.	91521	More Madrid intends to close the Isabel Zendal Pandemic Hospital if it wins the elections	0.1
7461	This is happening in Germany. People line up to stock up because of the coronavirus.	52456	In Germany, people line up outside this supermarket to stock up because of the coronavirus	0.2
769	official data of the uk show an increase of 5,400% in the number of women who have lost her baby after receiving COVID vaccines July 13, 2021	64587	The number of abortions in women who were vaccinated against COVID-19 in the United Kingdom has increased by 5,400%	0.17

Table 5: Examples of pairs that are not mapped as valid in the dataset. Cosine distance was computed using sentence-transformers/all-MiniLM-L6-v2.