NCL-AR at SemEval-2025 Task 7: A Sieve Filtering Approach to Refute the Misinformation within Harmful Social Media Posts

Alex Robertson
Newcastle University
Newcastle Upon Tyne, England
a.robertson4@ncl.ac.uk

Huizhi Liang
Newcastle University
Newcastle Upon Tyne, England
huizhi.liang@ncl.ac.uk

Abstract

With the overwhelming amount of online information and material, it is difficult to tell the truth from the lies. This poses a significant issue as individuals exploit this uncertainty to persuade, polarise, and misinform populations. However, multilingual fact-checking could be the solution by relating fake news to reliable sources worldwide, which is the main aim of SemEval-2025 Task 7. Our approach to this task employs an efficient sieve filtering method to retrieve the most relevant fact-checks that align with the social media post's original language, semantic content and uploaded time frame. Our approach achieved a success rate of 0.883 for the monolingual and 0.391 for the crosslingual tasks, while maintaining the high efficiency needed to limit the global impact of misinformation.

1 Introduction

Societally, we have been transitioning into a world of misinformation, disinformation, and fake news (Kavanagh and Rich, 2018). However, this transition has accelerated dramatically due to the recent advancement in generative AI, making it easier and more accessible to create highly believable and persuasive text, images, and videos (Nazar and Bustam, 2020; Xiong et al., 2024). The ability to post online as a form of expression, collaboration, and connection worldwide is becoming increasingly tainted by the spread of false and misleading information. This issue is becoming increasingly prevalent as widespread misinterpretations of these posts are fuelling polarisation, inciting hatred, and causing harm to the population (Broda and Strömbäck, 2024; Wu et al., 2019). Within this problem arises an essential need for fact-checking systems to combat this new threat. These systems must be highly accurate to counteract the false information and avoid amplifying the current issue. In addition, they must be efficient and multilingual, as around 5.22 billion social media users worldwide post and share content daily in various languages. The systems must identify and mitigate misinformation in real-time before it spreads (Kemp, 2024).

In the context of previously fact-checked claim retrieval (PFCR) (Shaar et al., 2020), Task 2 of CLEF 2022 focused on competitors determining a list of top-n fact-checked claims to corresponding tweets or political dialogues (Nakov et al., 2022). One of the published methods from the task, which achieved the highest accuracy, used a 3-step pipeline of pre-processing the tweets for clarity, retrieving the relevant claims using the BM25 model, and generatively re-ranking the claims using GPT-Neo-1.3B (Shliselberg and Dori-Hacohen, 2022). Another notable approach used the sentence transformer models, All-MiniLM-L6-v2 and All-MPNet-Base-v2, alongside a Support Machine Vector model to create an ensemble classifier to calculate the similarity score between tweets and fact-checked claims (Frick and Vogel, 2022).

Task 7 of SemEval-2025 aimed to address the challenge of misinformation (Peng et al., 2025). Participants were tasked with developing a fact retrieval system capable of providing relevant facts to contextualise or disprove social media posts. In this paper, we propose a sieve filtering-based approach that can retrieve facts to invalidate claims made in social media posts. The fact filters are initially coarse-grained, based on the original language of the social media posts, and end with fine-grained filters based on the exact time frame in which the posts were uploaded online. This streamlined approach achieved a 0.883 retrieval success rate in the monolingual task while maintaining a scalable efficiency level of processing a social media post per 0.07 seconds.

2 Dataset

The organisers used a modified version of the MultiClaim dataset, enhanced especially for the task

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 308–313
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

(Pikuliak et al., 2023). The authors provided a development dataset consisting of a post CSV P and fact-check CSV F, containing 24,431 posts and 153,743 fact-checks. In the testing dataset, this changed, with the post CSV changing to 8,276 unseen posts and the fact-check CVS expanding with an additional 118,704 unseen fact-checks, bringing the total number of fact-checks to 272,447. Each social media post within the dataset was flagged by a professional fact-checker from the social media sites Facebook, Instagram and Twitter. Each post entry $\mathcal{P}$ includes the post ID p_{id}, post Unix timestamp p_{time}, social media platform p_{plat}, post verdict p_{ver}, post content $p_{context}$ and OCR-extracted text p_{ocr}, which both contain the text in the original language, English translation, and predicted language country code with the accuracy rating.

$$\mathcal{P}_i = (p_{id}, p_{time}, p_{plat}, p_{ver}, p_{context}, p_{ocr})$$

$$P = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \ldots, \mathcal{P}_n\}$$

Each fact-check within the dataset was listed in the Google Fact Check Explorer or found in other sources such as Snopes. Each fact-check entry $\mathcal{F}$ includes the fact-check ID f_{id}, fact-check Unix timestamp f_{time}, fact-check article URL f_{url}, the fact-check article title f_{title} and claim f_{claim}, both structured identically to post text $p_{context}$.

$$\mathcal{F}_i = (f_{id}, f_{time}, f_{url}, f_{title}, f_{claim})$$

$$F = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \ldots, \mathcal{F}_n\}$$

3 Methodology

To tackle the task, we propose a simplistic sieve approach, as illustrated in Figure 1, to filter the $\mathcal{F}$ based on the post's features $(p_{time}, p_{context}, p_{ocr})$ until the top 10 most relevant $\mathcal{F}$ to the $\mathcal{P}$ are identified.

3.1 Fact Pre-Processing

Since we opted against using a machine learning approach for this task, pre-processing for all $\mathcal{F} \in F$ was a vital step to ensure high levels of accuracy. As mentioned in section 2, each $\mathcal{F}$ contained f_{title} and f_{claim} which held the main information regarding the fact. For the majority of $\mathcal{F}$, we would join the English translation of the texts to create f_{text} = "f_{claim}. f_{title}". However, in rare cases when f_{title} in $\mathcal{F}$ was empty, f_{text} = "f_{claim}". This concatenation maximises the semantic context in f_{text} as this will increase the accuracy during the similarity search. Then, given that the language style used within f_{text} was formal, $p_{context}$ contained slang, emojis, and grammatically incorrect punctuation. We utilised the regular expressions module to remove all non-alphabetical characters from f_{text}. This process will also be applied to $p_{context}$ later in the pipeline to align the textual styles. The final stage of the fact pre-processing was to cluster f_{text} based on the predicted language country code f_{lang} nested in f_{claim}, as shown by Figure 2. Preliminary experiments suggested that grouping the f_{text} based on f_{lang} significantly improved the overall accuracy and efficiency of the system, as this process would go on to reduce the size of the search space dramatically. The output of the fact pre-processing stage was a dictionary D containing clusters of f_{text}, grouped by their corresponding f_{lang}.

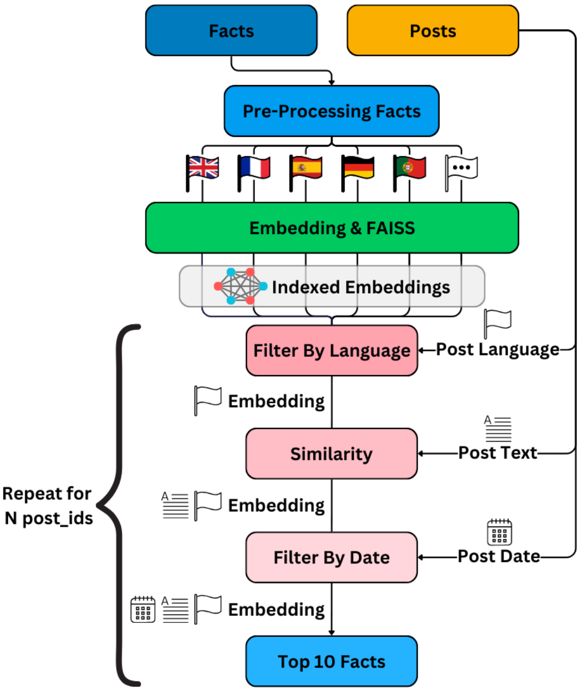

Figure 1: The sieve filtering approach

3.2 Embedding and FAISS Indexing

Unlike humans, computers struggle to compare the similarities of two text items. Therefore, the texts must be converted into numerical representations called embeddings. We used the Jina-Embedding-V3 for the embedding model (Sturua et al., 2024). This choice balanced accuracy and efficiency, achieving a score of 85.80 in the Sentence Textual Similarity (STS) section of the MTEB Leaderboard while maintaining a suitable size of

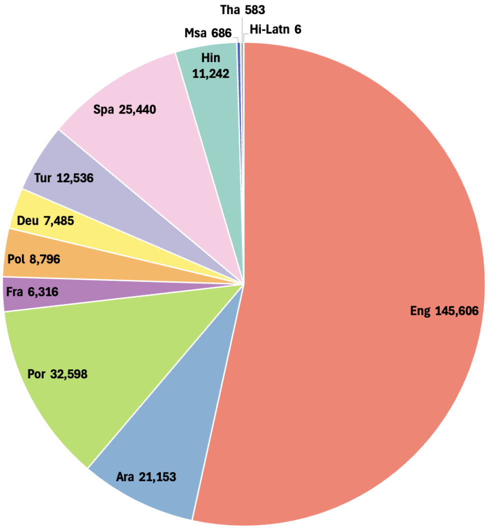

Figure 2: Pie chart of the language distribution for all $f_{lang} \in F$

570 million parameters (Muennighoff et al., 2023). We used the Flat Index L2 from the Facebook AI Similarity Search (FAISS) library to index and store the embeddings (Douze et al., 2024). The Flat Index L2 sequentially stores the embeddings and performs an exhaustive Euclidean distance search to find the most similar embeddings for a given query. This search approach is highly accurate, finding exact matches rather than approximations. However, this approach becomes inefficient and time-consuming for large datasets due to the sequential storing and brute-force search. This potential issue is mitigated by clustering the $\mathcal{F}$ based on f_{lang}, which partitions F into usable and efficient subsets. This is visualised in Figure 2, as regardless of $\mathcal{P}_{lang}$, the maximum dataset size used is reduced from 272,447 to 145,606 f_{text}.

$$\text{for all } L \in f_{\text{lang}}, \quad \text{for all } f_{\text{text}} \in D(L)$$

Each f_{text} is embedded into a 1024-dimensional vector $\mathbf{F}$ and stored in the FAISS Index $I(L)$.

3.3 Fact Sieving

The sieving stage of the approach applies a series of filters, each becoming more fine-grained, to $\mathcal{F}$ to identify the 10 most aligned $\mathcal{F}$ to each $\mathcal{P}$. The first layer filtered each $\mathcal{F}$ by the original language of $\mathcal{P}$, p_{lang}. We extracted the nested p_{lang} from $p_{context}$ if available; else, it was obtained from p_{ocr}. The extracted language was then used to retrieve the relevant facts from I based on p_{lang}, denoted as $I(p_{lang})$. On average, this simple filter reduced the number of $\mathcal{F}$ by 91.67% from 272,447 to 22,703, which boosted the performance accuracy by lowering the irrelevant $\mathcal{F}$ in the similarity search of the next layer. The second layer filtered each $\mathcal{F}$ by comparing the semantic content of f_{text} with the semantic content of $p_{content}$ and p_{ocr}. We concatenated the English translation of "$p_{content} \cdot p_{ocr}$" = p_{text}, which went through the same syntactical alignment pre-processing, as in section 3.1. We embedded p_{text} using Jina-Embedding-V3 to generate vector $\mathbf{P}$ and queried $I(p_{lang})$, which utilised Euclidean distance similarity search and retrieved the indexes of the top 250 most similar $\mathcal{F}$ to the $\mathcal{P}$ in $I(p_{lang})$. We converted the indexes to the evaluation output format of f_{id}.

$$S = \text{argsort}_{250}\left(-\|\mathbf{P} - \mathbf{F}_i\|_2\right), \\ \text{for all } \mathbf{F}_i \in I(p_{lang})$$

This layer is the most significant in terms of accuracy, as on average, it reduced the possible $\mathcal{F}$ by 98.80% from 22,703 to 250. We selected the top 250 as it balanced accuracy and efficiency while trying to reduce the inefficient, repetitive nature of the following filter. The third and final layer used the p_{time} to remove the outdated $\mathcal{F}$ according to their f_{time}. We set the cut-off time frame to approximately 9 months, as empirical testing indicated it was the optimum value to enhance accuracy without removing possibly relevant $\mathcal{F}$. As the majority of the timestamps were in the UNIX timestamp format, the $\mathcal{F}$ would be deemed outdated if:

$$|f_{\text{time}} - p_{\text{time}}| > 9 \times 30 \times 24 \times 60 \times 60$$

We would iterate through the S, removing the outdated $\mathcal{F}$ from the similarity-ordered list using their corresponding f_{id}. The last step was to choose the first 10 items from S, which reduced the number of $\mathcal{F}$ by 96% from 250 to the final 10 f_{id} needed for Task 7 of SemEval-2025.

4 Evaluation Setup

The evaluation of SemEval-2025 Task 7 was split into two tracks: monolingual and crosslingual. The claim and the retrieved fact-check were in the same language in the monolingual track. Whereas the crosslingual track meant the claim and the retrieved fact-check may be in different languages from the 27 available in the test dataset, which provided an extra dimension of complexity to the challenge.

4.1 Evaluation Data

For the evaluation, alongside the P and F, the authors also provide 2 JSON submission files for the

Name	Size (M)	Mono (S@10)											Cross (S@10)
		avg	eng	fra	deu	por	spa	tha	msa	ara	tur	pol	**avg**
Multilingual-e5-Base	278	0.731	0.778	0.726	0.758	0.722	0.754	0.672	0.709	0.874	0.670	0.648	0.704
Bilingual-Embedding-Base	278	0.738	0.776	0.738	0.766	0.724	0.796	0.705	0.634	0.890	0.684	0.666	0.715
GTE-Large-en-1.5v	434	0.669	0.724	0.690	0.684	0.614	0.704	0.661	0.656	0.794	0.546	0.618	0.617
Jina-Embedding-V3	572	0.751	0.758	0.770	0.768	0.710	0.796	0.721	0.666	0.902	0.716	0.702	0.685
Approach$_{\text{Submitted}}$		0.872	0.832	0.918	0.898	0.798	0.878	0.945	0.946	0.884	0.814	0.800	0.398
Approach$_{\text{Best}}$		0.883	0.816	0.928	0.882	0.782	0.876	0.973	0.989	0.934	0.828	0.820	0.391
Leaderboard$_{\text{Best}}$		**0.960**	**0.916**	**0.972**	**0.958**	**0.926**	**0.974**	**0.995**	**1**	**0.986**	**0.948**	**0.926**	**0.859**

Table 1: Comparison of Performance using the S@10 scores among the baseline textual embedding models (Multilingual-e5-Base (Wang et al., 2024), Bilingual-Embedding-Base (Lajavaness, 2024), GTE-Large-en-1.5v (Li et al., 2023), and Jina-Embedding-V3), our approaches, and the leading method of SemEval-2025 Task 7

monolingual and crosslingual tracks. These files contain a respective 4276 and 4000 p_{id} corresponding to $\mathcal{P}$, each with an empty list to be filled with the 10 most relevant f_{id}, as described in 3.3.

4.2 Evaluation Metrics

The performance of the approaches in Task 7 are measured using the Success-at-10 (S@10). Each $\mathcal{P}$ has one or more corresponding $\mathcal{F}$. A retrieval is successful if the golden $\mathcal{F}$ is within the 10 submitted $\mathcal{F}$. The success-at-10 rate is average over the number of p_{id} entries in the JSON file. The monolingual evaluation goes slightly further than the crosslingual evaluation, as the S@10 is additionally calculated for each language for a more in-depth analysis.

5 Results

5.1 Main Results

The results in Table 1 indicate that Approach$_{\text{Best}}$ achieved our highest retrieval success rate with 0.883 in the test-phase monolingual task. This suggests our approach can accurately provide the factual information needed to disprove the misinformation in potentially harmful social media posts. The filtering rules of our approach were consolidated during the development phase of SemEval-2025. For a comparison, Approach$_{\text{Best}}$ was evaluated post-SemEval on the development-phase monolingual task, it scored a retrieval success rate of 0.832. The improvement when faced with new data in the test phase shows the generalizability of the filtering rules of our approach. In the more complex crosslingual task, we achieved an accuracy of 0.391. This lower performance can be attributed to our approach not aligning with the cross-lingual task, as we reduce the search space by focusing on $\mathcal{F}$ where $\mathcal{F}_{lang}$ is the same as $\mathcal{P}_{lang}$. However, this is counterintuitive, as for the cross-lingual task, we should prioritise $\mathcal{F}$ and $\mathcal{P}$ with different languages. Addressing this misalignment could make the cross-lingual approach equally as accurate as the monolingual task. In SemEval-2025 Task 7, the highest recorded scores for the monolingual and crosslingual tasks were a remarkable 0.960 and 0.859, respectively. While our approach has lower accuracy, its greater efficiency compared to smaller text embedding models, as shown in Table 2, makes it a strong baseline approach that can be built upon and further refined.

5.2 Efficiency

In addition to aiming for high multilingual accuracy, another main goal of the approach was high efficiency, as this would ensure scalability in handling the volume of social media posts online. Each text embedding model and our strategies were timed from the initialisation of the script through the embedding phase until the final $\mathcal{P}$ had been countered by $\mathcal{F}$. The Jina-Embedding-V3 model had the quickest processing time amongst the baselines. Despite its larger size of 572 million parameters, it was over a minute and 20 seconds quicker than Multilingual-e5-Base, which is half the size. However, our best approach outperformed all baselines in both tasks, reducing the time taken by around 3 minutes from Jina-Embedding-V3. This significant gain in efficiency is due to language filtering, which reduces the size of search space, making the $\mathcal{F}$ retrieval faster and more efficient.

5.3 Ablation Study

It was logical that a relevant pairing of $\mathcal{P}$ and $\mathcal{F}$, would be within a similar time frame. This assumption was validated while analysing and manually pairing $\mathcal{P}$ to $\mathcal{F}$ in the development dataset, as there was a strong correlation between the time frame and the content shared between similar $\mathcal{P}$ and $\mathcal{F}$. However, we acknowledged that a time filtering

Name	Eval Time (s)		Time Per P (s)	
	Mono	Cross	Mono	Cross
Multilingual-e5-Base	585	569	0.137	0.142
Bilingual-Embedding-Base	619	608	0.145	0.152
GTE-Large-en-1.5v	2134	2077	0.499	0.519
Jina-Embedding-V3	505	480	0.118	0.120
Approach$_{\text{Submitted}}$	3140	3270	0.734	0.818
Approach$_{\text{Best}}$	**304**	**312**	**0.07**	**0.08**

Table 2: Comparison of Efficiency using the evaluation completion time among the baseline textual embedding models and our approaches

step would be the most inefficient part of the sieving process due to its repetitive nature. Nonetheless, we decided to prioritise accuracy over efficiency. The Approach$_{\text{Submitted}}$ includes the time filtering step, and Approach$_{\text{Best}}$ does not. As evident in Table 1 and 2, the previously mentioned assumption was incorrect, as Approach$_{\text{Best}}$ has marginally higher accuracy by 0.03 while being significantly more efficient by over an hour and a half across the two tasks. Unfortunately, we could not recognise this mistake before the SemEval-2025 Task 7 test-phase ended.

6 Conclusion

In this paper, we presented our approach in the SemEval-2025 Task 7, which addressed the ongoing challenge of fact-checking social media posts to debunk the misinformation contained within. The task has an additional difficulty component, as the facts and posts are from one of 27 different languages, reflecting the global importance of this challenge. Our sieve filtering approach used the key features, such as original language, semantic content, and time frame, from each social media post to retrieve the 10 most relevant fact-checks. Our approach achieved a 0.883 success rate when retrieving fact-post parings in the monolingual task without sacrificing the high-efficiency levels of 0.07 seconds to disprove each post factually. While our approach does not match the top approaches in the leaderboard, it provides a strong baseline for future improvements. One future improvement for the approach would be to use the other features within each fact-check, such as the article URL, to provide more context and potentially improve retrieval accuracy.

References

Elena Broda and Jesper Strömbäck. 2024. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2):139–166.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Raphael Antonius Frick and Inna Vogel. 2022. Fraunhofer sit at checkthat!-2022: Ensemble similarity estimation for finding previously fact-checked claims. In *Conference and Labs of the Evaluation Forum*.

Jennifer Kavanagh and Michael D. Rich. 2018. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. RAND Corporation, Santa Monica, CA.

Simon Kemp. 2024. Datareportal – global digital insights.

Lajavaness. 2024. bilingual-embedding-base.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *Advances in Information Retrieval*, pages 416–428, Cham. Springer International Publishing.

S Nazar and M R Bustam. 2020. Artificial intelligence and new level of fake news. *IOP Conference Series: Materials Science and Engineering*, 879(1):012006.

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. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, Singapore. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Michael Shliselberg and Shiri Dori-Hacohen. 2022. Riet lab at checkthat!-2022: Improving decoder based re-ranking for claim matching. In *Conference and Labs of the Evaluation Forum*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90.

Feng Xiong, Thanet Markchom, Ziwei Zheng, Subin Jung, Varun Ojha, and Huizhi Liang. 2024. NCL-UoR at SemEval-2024 task 8: Fine-tuning large language models for multigenerator, multidomain, and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 163–169, Mexico City, Mexico. Association for Computational Linguistics.