RAG and Recall: Multilingual Hate Speech Detection with Semantic Memory

Khouloud Mnassri, Reza Farahbakhsh and Noel Crespi
Samovar, Télécom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France.
khouloud.mnassri@telecom-sudparis.eu

Abstract

Multilingual hate speech detection presents a challenging task, particularly in limited-resource contexts when performance is affected by cultural nuances and data scarcity. Fine-tuned models are often unable to generalize beyond their training, which limits their efficiency, especially for low-resource languages. In this paper, we introduce HS-RAG, a retrieval-augmented generation (RAG) system that directly leverages knowledge, in English, French, and Arabic, from Hate Speech Superset (publicly available dataset) and Wikipedia to Large Language Models (LLMs). To further enhance robustness, we introduce HS-MemRAG, a memory-augmented extension that integrates a semantic cache. This model reduces redundant retrieval while improving contextual relevance and hate speech detection among the three languages.
Warning: This document contains some examples of hateful content.

1 Introduction

Hate speech is one of the most common categories of online abuse and harm. Its detection is a vital task in natural language processing (NLP), in order to ensure safe online communication. Nevertheless, it remains challenging due to its coded language and cultural nuance (Vidgen et al., 2019; Mnassri et al., 2024). Therefore, detecting this content goes beyond the need for accurate classification to requires cultural and linguistic adaptability. On the other hand, the multilingual aspect of online abuse and harms has gained more awareness as online platforms are progressively serving more global audiences. As a result, recent studies about hate speech have been focusing more on multilingualism (Mnassri et al., 2024), with growing attention to reducing the English bias in this field of study (Tonneau et al., 2024).

Recently, Large Language Models (LLMs) have revolutionized NLP tasks (Brown and Mann, 2020). More specifically, they have efficiently contributed to hate speech detection, particularly in multilingual and cross-cultural settings (Albladi et al., 2025). However, LLMs still encounter significant challenges in terms of resource consumption (Bai et al., 2024a), computational inefficiency (Bai et al., 2024b), hallucination (Liu et al., 2022), and misfollowing instructions (Ouyang et al., 2022).

An efficient alternative approach introduced by (Lewis et al., 2020) is Retrieval-Augmented Generation (RAG), which provides promising solutions to these issues (Izacard et al., 2023). By combining the strength of LLMs with knowledge retrieval techniques, RAG enables models to enhance their understanding by adding relevant information from external data sources (Gao et al., 2023), which helps to outperform fine-tuned LLMs (Chen et al., 2024).

Nevertheless, RAG-based systems often encounter some difficulties, such as high retrieval latency, as they continually query large databases. This redundant retrieval procedure not only slows down inference, but can also

degrade the quality of generated outputs, especially in real-time applications (Chan et al., 2024). To address these challenges, different caching mechanisms have been examined (Jin et al., 2024). One of the caching techniques is semantic caching, which enables knowledge reuse based on embedding similarity. This approach is relevant for multilingual NLP tasks, where the same data sample can be represented differently across languages. This explains the tendency to use this caching mechanism in LLMs same as in GPTCache (Bang, 2023).

In this paper, we present HS-RAG and HS-

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 219–227
August 1, 2025 ©2025 Association for Computational Linguistics

MemRAG, two novel architectures for multilingual hate speech detection and moderation of online abuse and harms. HS-RAG implements a multilingual RAG pipeline, incorporating Wikipedia and hate speech datasets for context-aware detection. HS-MemRAG extends this with a semantic cache memory, reducing redundant retrievals and reusing contextual similarity. Our models offer robust performance across languages, providing an explainable, lightweight, and adaptable solution for multilingual content moderation especially in low-resource languages.

2 Methodology

2.1 Data

Structured data - Hate Speech Superset (HS dataset): Proposed by Tonneau et al. (Tonneau et al., 2024), this dataset is an open access multilingual corpus[1]. Due to computational constraints, we focus on three languages: English (En), French (Fr), and Arabic (Ar). These were selected based on their linguistic diversity, and our own linguistic expertise. In order to get a balance between languages, we downsampled En and Ar datasets into 18.000 random samples (to get the same size as Fr dataset). After concatenation, we got a final multilingual corpus we used for retrieving and fine-tuning. As for testing, we randomly selected 1000 samples per language.

Non structured data - Wikipedia: We used Wikipedia in our RAG-based models for better contextual understanding and generation. By setting a maximum number of 100 documents per keyword, with 1000 characters per document, we managed to extract 1093 documents by automatically searching for their titles based on specific keywords: 'Hate speech','Offensive languag','Cyberbullying' and 'Hate crime', for En, same translated expressions are used to extract in Fr and Ar.

2.2 HS-Base - Zero-shot & Fine-tuning

As baselines, we build Hate Speech HS-Base, a multilingual hate speech classifier using Meta-LLaMA-3-8B model (AI@Meta, 2024). We experiment with both zero-shot and fine-tuned variants to assess the model's ability to detect hate speech. These baselines help us to evaluate the intrinsic multilingual capabilities of the LLM, and to understand the gains and limitations of fine-tuning compared to retrieval-augmented approaches.

To get a good performance while lowering resource consumption, we employed parameter-efficient fine-tuning (PEFT), more specifically LoRA (Low-Rank Adaptation) (Hu et al., 2021), and 4-bit quantization. In order to avoid any class imbalances, class weights were also calculated.

We added sequence classification heads to shape the HS-Base models as classifiers in order to only output the required labels among the two categories 'hateful' and 'non-hateful'.

We utilize both HS dataset and Wikipedia data in order to mimic realistic settings, where the top-k retrieved documents are usually expected to be from different sources. This provides a real-world scenario where prepared, labeled datasets are unavailable or scarce. Unlike fine-tuning, our RAG-based models retrieve context without parameter updates, enabling training-free, and multilingual deployment with more flexibility and robustness.

2.3 HS-RAG

To ensure that responses are accurate, contextually aware, and less confronted to hallucinations, we built HS-RAG, which makes use of a retrieval mechanism to improve the generation process for final multilingual hate speech detection.

The main elements of our models are generator and retrieval. To predict the proper label for a given data sample d in language L_d, these components smoothly incorporate retrieval-augmented techniques with deep natural language understanding. The overall structure of our model is presented in Figure 1, which is composed of:

Multilingual Hate Speech Retrieval: The retrieval part searches relevant contextual information from $Langchain$ vector database, $Chroma$ vector store[2], referenced as C. The retrieving process is defined as $c = R(d, C)$, where d indicates the data sample, and R accepts the top K relevant documents: $c = topK_{t \in C}(h_d^T, h_{doc})$, ranking the similarity scores (cosine similarity) between data samples embeddings h_d^T and document embeddings h_{doc}.

Multilingual Hate Speech Generation & Mapping: The relevant data, obtained after retrieving context $R(d, K)$, is passed to a pre-trained LLM for generating predicted $answer$ (where

[1] https://huggingface.co/collections/manueltonneau/hate-speech-supersets-664ef6d2bc40cce7a8b1092f

[2] https://python.langchain.com/docs/integrations /vectorstores/chroma/

we set $max\_new\_tokens = 200$): $answer = f_{LLM}R(d, K)$. Then, $answer$ is mapped to get the final label: ‘hateful’ or ‘non-hateful’.

We expected that the LLM might not answer or generate unexpected outputs. Therefore, our mapping function checks for empty or malformed responses and assigns them with a fallback value (-1). Nevertheless, in practice, our models gave usable outputs so we didn’t encounter any fallbacks during evaluation.

For the generation process, we employ a prompt template,

which instructs a multilingual hate speech expert to detect hate speech of a given text in (En, Fr, or Ar), based on the retrieved *context*.

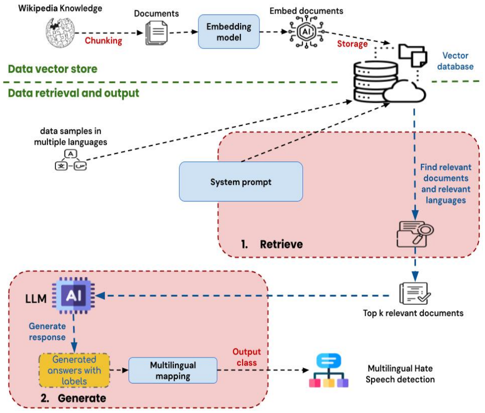


Figure 1: HS-RAG structure.

2.4 HS-MemRAG - Semantic Cache Memory

As displayed in Figure 2, we have integrated semantic cache into the retrieval part of HS-RAG model. This caching intercepts data samples before proceeding to the main vector database. More specifically, if a data sample is semantically similar to previously processed one, the latter’s cached response is directly returned. Thus, bypassing duplicative retrieval and computation.

The semantic cache performs by calculating semantic similarity utilizing dense vector embeddings. Using pre-trained embedding model, it develops embeddings for every data sample and stores their responses. To determine similarity, the system compares embeddings using Cosine similarity. If a new sample is close enough to an existing one (within a predefined threshold), the corresponding stored response is returned directly from the cache, thus, avoiding retrieving from the vector database (Jin et al., 2024). This approach helps reduce repetitive retrievals, improving efficiency without compromising retrieval quality.

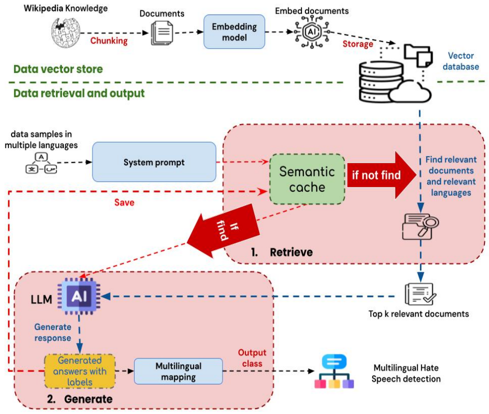


Figure 2: HS-MemRAG structure.

3 Experiments And Results

We present experimental details in Appendix A. In Appendix B, we describe the positioning of a data sample among the data stored in our $ChromaDB$, in order to better understand the retrieval process. Table 1 illustrates the evaluation results of our models focusing on four key performance metrics: Accuracy (Acc.), Precision (Pr.), Recall (Rc.), and weighted F1-score (F1). We also highlight values that indicate the highest performance results in bold between HS-Base models, and between HS-RAG and HS-MemRAG models.

Model	Lang.	Acc.	Pr.	Rc.	F1
HS-BASE (Zero-shot)	English	0.657	0.627	0.657	0.6398
	French	0.5815	0.6175	0.5815	0.5975
	Arabic	0.3265	**0.9255**	0.3265	0.4533
HS-BASE (Fine-tuned)	English	0.7515	0.8079	0.7515	0.7653
	French	0.7955	0.8301	0.7955	0.8056
	Arabic	**0.959**	0.9206	**0.959**	**0.9394**
HS-RAG	English	0.67	0.7735	0.67	0.692
	French	0.759	0.7652	0.759	0.7619
	Arabic	0.859	0.9345	0.859	0.8918
HS-MemRAG	English	0.668	0.7715	0.668	0.6902
	French	0.702	0.7082	0.702	0.705
	Arabic	**0.888**	**0.9381**	**0.888**	**0.9099**

Table 1: Performance across models and languages.

Results in Table 1 are based on a single run. In fact, rerunning the models consistently gave similar results, illustrating that the outputs are stable, with minimal variance expected.

Our results indicate that while fine-tuning gives the highest overall performance, both HS-RAG and HS-MemRAG propose effective alternatives, especially when considering multilingual generalization. HS-RAG enhances over zero-shot HS-Base baseline across all three languages, showing that retrieval-augmented data enables bridging knowledge gaps, particularly in Fr (+16 F1) and Ar (+44 F1). These improvements are mostly important in limited-resource environments, where training data is scarce and computational resources are constrained. They confirm the expected advantage of retrieval based models in using external knowledge to improve detection performance.

Moreover, HS-MemRAG demonstrates its most significant effeciency in Ar, where it performs the highest F1 score overall (0.91). While the improvements over HS-RAG in En and Fr are modest, HS-MemRAG presents clear efficiency usefulness by reducing redundant retrieval on semantically similar inputs, thus enabling faster inference.

Although neither RAG nor MemRAG surpassed the fine-tuned model in most cases, they did offer strong zero-training performance, and easy deployment in multilingual settings. This makes them attractive options for real-world scenarios where compute, structured data, or time is constrained.

While fine-tuned HS-BASE model gave the highest F1-score in Arabic (0.9394), its confusion matrix indicates that it fails to detect hateful samples. It achieves an F1-score of 0.0 for this label, and predicting almost all test samples as non-hate. This result illustrates the impact of class imbalance in inflating performance metrics. Nevertheless, both HS-RAG and HS-MemRAG prove their ability to detect hate speech, achieving F1-scores of 0.18 and 0.23 respectively for this minority class. This confirms that, unlike fine-tuning, our models do not collapse into majority class predictions and demonstrate stronger generalization in low-resource, imbalanced settings.

Overall, we could understand why Arabic performance may seem more robust than English in some settings, since metrics can be skewed by model behavior toward the majority class.

3.1 Retrieved Languages in HS-RAG

To further understand how our retrieval approach is executed across languages, we examine the languages of the documents retrieved by HS-RAG model during inference. In fact, for each data sample of the test sets, we were tracking the language metadata of the top-*k* retrieved documents and visualizing their frequency distribution. Table 2 displays the distribution of retrieved document languages for every test set.

As shown in Table 2, we demonstrate that HS-RAG often retrieves documents from different languages, depending mainly on context similarity rather than language similarity. In fact, for Arabic test set, most retrieved documents were in Arabic and French, with almost no English retrieving, suggesting strong cross-lingual similarity between Arabic and French in hate speech. As for French test set, HS-RAG mostly retrieved French documents with a noteworthy number from Arabic, showing bidirectional overlap. On the other hand, English inputs surprisingly retrieved more French than English documents, implying that French hateful data may provide more semantically aligned hate speech content in certain English contexts.

Test Set	Retrieved En	Retrieved Fr	Retrieved Ar
English	545	**828**	527
French	28	**986**	333
Arabic	12	398	**959**

Table 2: Languages of documents retrieved.

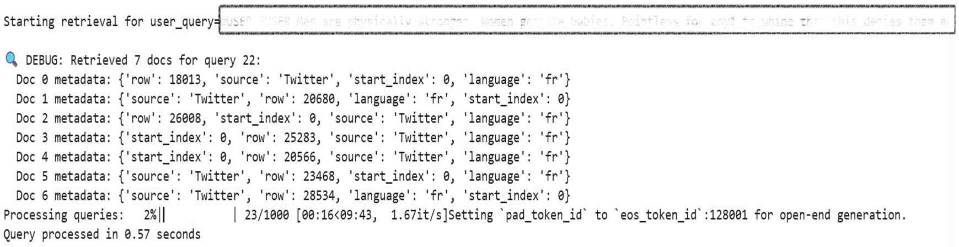
For example, in an English test sample, HS-RAG retrieved the top-7 documents, all in French (Figure 3). We believe that this behavior is because certain topics can be expressed more explicitly in the French dataset we used, which results in stronger semantic matches in the embedding space. Overall, this ability to retrieve semantically relevant context across different languages contributes to HS-RAG detection performance, especially in low-resource or ambiguous inputs.

In general, these results emphasize that cross-lingual retrieval enables generalization in multilingual hate speech. Therefore, the HS-RAG model presents a key advantage for multilingual environments, leveraging shared hateful content across languages. Moreover, although retrieving relevant documents in a different language may introduce noise, this issue is mitigated by employing multilingual sentence embeddings A.2, which capture deep semantic patterns across languages.

3.2 Retrieval source in HS-MemRAG: From Vector Database or from Cache Memory

To understand the influence of semantic caching in HS-MemRAG, we measure the frequency of the retrieval source used for each test dataset: for each data sample, we track whether the result came from the cache memory or from the vector database (DB). Results are displayed in Table 3.

For each language, we also report the detection accuracy per source by measuring the percentage



```
Starting retrieval for user_query=[redacted]

🔍 DEBUG: Retrieved 7 docs for query 22:
  Doc 0 metadata: {'row': 18013, 'source': 'Twitter', 'start_index': 0, 'language': 'fr'}
  Doc 1 metadata: {'source': 'Twitter', 'row': 20680, 'language': 'fr', 'start_index': 0}
  Doc 2 metadata: {'row': 26008, 'start_index': 0, 'source': 'Twitter', 'language': 'fr'}
  Doc 3 metadata: {'start_index': 0, 'row': 25283, 'source': 'Twitter', 'language': 'fr'}
  Doc 4 metadata: {'start_index': 0, 'row': 20566, 'source': 'Twitter', 'language': 'fr'}
  Doc 5 metadata: {'source': 'Twitter', 'row': 23468, 'start_index': 0, 'language': 'fr'}
  Doc 6 metadata: {'source': 'Twitter', 'row': 28534, 'language': 'fr', 'start_index': 0}
Processing queries:   2%||          | 23/1000 [00:16<09:43,  1.67it/s]Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Query processed in 0.57 seconds
```

Figure 3: Example of retrieved documents for an English query - Dominance of French documents retrieved. The query content is partially blurred in accordance with WOAH's reporting policy on abusive language.

Language	Cache Usage (%)	Vector DB Usage (%)
English	1.0	99.0
French	28.1	71.9
Arabic	30.6	69.4

Table 3: Retrieval source usage per language.

of correct predictions when each source was used alone. Results are presented in Table 4.

Language	Cache Accuracy	Vector DB Accuracy
English	**0.70**	0.33
French	**0.48**	0.23
Arabic	0.08	**0.13**

Table 4: Accuracy by retrieval source per language.

We observe that cache memory usage differs significantly across languages, it is scarcely used in English (1%) but widely utilized in French (28%) and Arabic (31%). In addition, the cache memory provides higher accuracy than vector database retrieval in English and French, enabling faster and more stable prediction. Based on Table 4, the semantic cache provides more accurate retrieval matches in English and French. However, in Arabic, the cache didn't perform well than vector DB retrieval. This can be related to the high linguistic variety in Arabic hate speech expressions, especially that it has several dialects. Therefore, vector DB retrieval presents better adaptation.

4 Related Work

4.1 RAG for Multilingual Hate Speech detection

Despite its significant exploration in several domains like knowledge-intensive tasks and question-answering (Lewis et al., 2020; Yu, 2022; Cai et al., 2022), RAG's use in classification, particularly in hate speech detection, is still unexplored. We found two studies proposing RAG-based models as counter hate speech generators (Jiang et al., 2023; Leekha et al., 2024). The use of RAG in multilingual aspect is still also in its beginning (Gao et al., 2022; Wang et al., 2023; Chirkova et al., 2024), we found a study (Yao et al., 2024) investigating cross-cultural moderation using RAG in Korean.

4.2 First Steps in Memory Caching for Hate Speech Moderation

Starting with LLMs, GPTCache facilitates retrieval using semantic similarity (Bang, 2023). Adding to that, (Gill et al., 2024) presented cache based on Federated Learning. Moreover, (Li et al., 2024) introduced a cache with optimized storage strategies. Also, (Mohandoss, 2024) proposed a context-based semantic cache leveraging query context.

As for RAG-based approaches, (Jin et al., 2024) addressed long sequences through caching. Besides that, (Lu et al., 2024) presented a key-value based cache.

Despite the increased interest in retrieval-augmented and memory-based approaches, we found no previous study using memory caching for hate speech detection task. This indicates our contribution to a novel step towards robust online abuse and harms moderation systems.

5 Conclusion

We proposed HS-RAG and HS-MemRAG, two training-free, multilingual hate speech detection models that can manage moderation systems in a range of online abuse and harms detection tasks. They can be extended to other types of harms (e.g., misinformation, radicalization). Our models leverage retrieval augmentation and semantic cache memory, providing robust performance across English, French, and Arabic. By incorporating semantic cache, HS-MemRAG enables faster and more stable predictions, offering a lightweight and explainable solution for multilingual content moderation, especially for under-resourced languages (Arabic).

6 Limitations

Due to computational constraints, our experiments were limited to a subset of Hate Speech Superset dataset (Tonneau et al., 2024). We were restricted to only three languages, with 1000 test samples per language. As a result, we couldn't study our models' generalization capacity across more languages and bigger datasets. For more systematic multilingual insights, we aim to extend our analyses to the entire Hate Speech dataset in our future work, if resources allow.

Additionally, we discovered that, compared to the cache memory, the main vector database still provides the majority of the retrieval context in HS-MemRAG. We believe this is related to the long size of data samples and the difficulties with semantic similarity. Therefore, we seek to investigate more refined semantic caching strategies in future research to improve retrieval efficiency.

Moreover, thanks to its compatibility with our computational resources, we used LLaMA 3. However, we aim to explore other open-source LLMs (e.g., Mistral, Zephyr, Gemma, GPT) in the future, to better comprehend performance variability among different architectures.

References

AI@Meta. 2024. Llama 3 model card.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Xinyuan Song, Carl Yang, Yue Cheng, and Liang Zhao. 2024a. Beyond efficiency: A systematic survey of resource-efficient large language models. *Preprint*, arXiv:2401.00625.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Fu Bang. 2023. GPTCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 212–218, Singapore. Association for Computational Linguistics.

Tom Brown and Benjamin et al. Mann. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3417–3419, New York, NY, USA. Association for Computing Machinery.

Brian J Chan, Chao-Ting Chen, Jui-Hung Cheng, and Hen-Hsen Huang. 2024. Don't do rag: When cache-augmented generation is all you need for knowledge tasks. *Preprint*, arXiv:2412.15605.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, and Michael Lyu. 2022. Retrieval-augmented multilingual keyphrase generation with retriever-generator iterative training. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1233–1246, Seattle, United States. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Waris Gill, Mohamed Elidrisi, Pallavi Kalapatapu, Ammar Ahmed, Ali Anwar, and Muhammad Ali Gulzar.

2024. Meancache: User-centric semantic cache for large language model based web services. *Preprint*, arXiv:2403.02694.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. 2024. Ragcache: Efficient knowledge caching for retrieval-augmented generation. *Preprint*, arXiv:2404.12457.

Rohan Leekha, Olga Simek, and Charlie Dagli. 2024. War of words: Harnessing the potential of large language models and retrieval augmented generation to classify, counter and diffuse hate speech. *The International FLAIRS Conference Proceedings*, 37(1).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiaxing Li, Chi Xu, Feng Wang, Isaac M von Riedemann, Cong Zhang, and Jiangchuan Liu. 2024. Scalm: Towards semantic caching for automated chat services with large language models. *Preprint*, arXiv:2406.00025.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Songshuo Lu, Hua Wang, Yutian Rong, Zhi Chen, and Yaohua Tang. 2024. Turborag: Accelerating retrieval-augmented generation with precomputed kv caches for chunked text. *Preprint*, arXiv:2410.07590.

Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, Guanlin Li, and Noel Crespi. 2024. A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934.

Ramaswami Mohandoss. 2024. Context-based semantic caching for llm applications. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 371–376.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Red Hook, NY, USA. Curran Associates Inc.

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.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023. Retrieval-augmented multilingual knowledge editing. *arXiv preprint arXiv:2312.13040*.

Tsungcheng Yao, Ernest Foo, and Sebastian Binnewies. 2024. Personalised abusive language detection using LLMs and retrieval-augmented generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 92–98, Trento. Association for Computational Linguistics.

Wenhao Yu. 2022. Retrieval-augmented generation across heterogeneous knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 52–58, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

A Experimental settings

Experiments were executed on Google Colab Pro+ using NVIDIA A100 GPU.

A.1 HS-Base

Model	HuggingFace `Meta-Llama-3-8B`
Optimization	4-bit Quantization, LoRA (Low-Rank Adaptation)
Training Details	1 Epoch, Cross-Entropy Loss
Batch Size	32
Learning Rate	2e-5

Table 5: HS-Base parameters.

We implement the LLM model, here, as a classifier rather than generator, so we didn't use any prompting to build HS-Base. We instead compare against standard zero-shot (non trianed) and fine-tuned (trained) classifiers.

A.2 HS-RAG

1. $Wikipedia$ Python library to collect knowledge from Wikipedia.

2. `Document` class to store collected content and metadata (title, language, source).

3. Langchain $CSVLoader$ to load hate speech structured data and tag obtained documents with source information.

4. $RecursiveCharacterTextSplitter$ to split documents into chunks: 1500 characters and overlap of 300.

5. $xlm-roberta-base$ (Conneau et al., 2020) to tokenize chunks.

6. $paraphrase-multilingual-mpnet-base-v2$ (Reimers and Gurevych, 2019) sentence transformer to embed-chunks, and to store them in **Chroma** vector database.

7. Prompt to generate predicted answers based on input text and retrieved context.

8. RAG pipeline configuration to leverage stored embeddings for context retrieval and integrate a 4-bit quantized LLM for answer generation.

Table 6: HS-RAG Pipeline.

Tools: *Wikipedia*[3], Langchain *CSVLoader*[4], *Recursive Character Text Splitter*[5]

Below is the prompt we used:
'You are a multilingual hate speech detection expert in English, French, and Arabic. Your task is to classify the following text as Hateful or Non-hateful based on its content.
–Hateful: Clearly expresses hate, promotes violence, or dehumanizes people based on identity (e.g., race, religion, gender, nationality).
–Non-hateful: May express strong opinions, offense, or disagreement, but does not incite hate or target identity groups.
Instruction: Be strict, classify as Hateful only when there is a clear attack on an identity group. Respond with one word: Hateful or Non-hateful.'

A.3 HS-MemRAG

For each data sample, the semantic cache system first checks if a relevant response exists in the cache memory, based on a specific threshold. If no similarity is found, the data sample is forwarded to the vector database for document retrieval, and the cache is edited with the newly generated result. More details about the parameters are presented below.

Setting	**Role**
Embedder	$paraphrase-multilingual-mpnet-base-v2$ to generate multilingual sentence embeddings.
Index	Employs $Annoy$ with `angular` distance to approximate cosine similarity during nearest-neighbor search.
JSON File	Stores input samples along with their generated predictions for persistent caching.
Similarity Threshold	Set to 0.9. Cached responses are reused when angular distance between embeddings falls below this threshold.

Table 7: HS-MemRAG semantic cache settings.

[3] https://pypi.org/project/wikipedia/

[4] https://python.langchain.com/docs/integrations/document_loaders/csv/

[5] https://python.langchain.com/docs/how_to /recursive_text$_{s}plitter/$

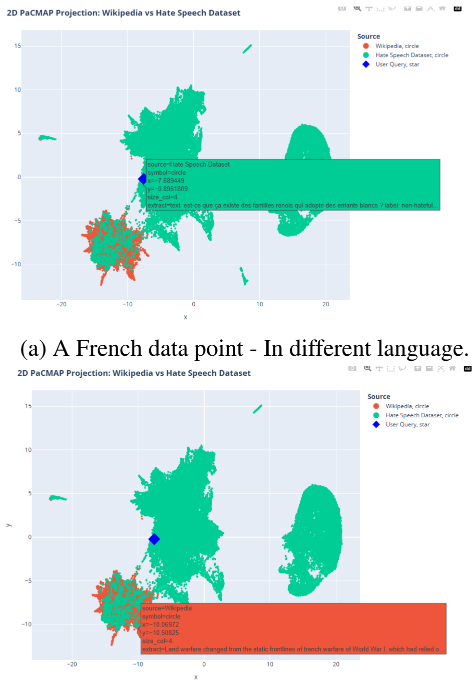
Tools: *Annoy*[6]

We tested different threshold values and found that 0.9 gave the best overall performance, providing a balance between avoiding redundant responses and retrieval precision.

B Semantic Proximity visualization for Cross-Domain Retrieval

Figures 4 display an example of data sample from the English test dataset and its positioning between the embeddings stored in our Chroma database .
The example we randomly select from the English test set is: *‘, no. We don’t need more N*****s to represent us. In the end they always work for their people just as much as we need to work for our people.’*
Using PaCMAP[7], we managed to see the position of this data sample among the data points of the two data sources: Hate Speech Dataset and Wikipedia.

Figures 4 indicate that the two data sources provide pertinent content for the data sample, but at various levels. For this example, the embeddings are placed significantly closer to samples from the Hate Speech dataset than to Wikipedia documents in the vector space. This indicates that the data example shares stronger contextual and linguistic similarity with the Hate Speech dataset more than with Wikipedia data.



(a) A French data point - In different language.

(b) An English data point - In same language.

Figure 4: Example of data point positioning in the vector database.

Also, although the majority of the closest data points were in the same language as the data example (in English), some others were in different languages but had the same context. Therefore, we demonstrate that the retrieval method tends to select semantic similarity and contextual relevance over linguistic overlap. More specifically, we observe that a different language data point (in French in Figure 4a) appears to be closer to the data sample than a same language one (in English in Figure 4b). This demonstrates that the retrieval method concentrates more on contextual similarities rather than linguistic similarities.

[6] https://pypi.org/project/annoy/
[7] https://github.com/YingfanWang/PaCMAP