Embedding The Truth: Approximate Caching for Fact Checking

Abdullah Ghani[*]
25100155@lums.edu.pk
Lahore University of Management Sciences (LUMS)
Lahore, Punjab, Pakistan

Danish Athar[*]
25100174@lums.edu.pk
Lahore University of Management Sciences (LUMS)
Lahore, Punjab, Pakistan

Muhammad Ayain Fida Rana[*]
25100045@lums.edu.pk
Lahore University of Management Sciences (LUMS)
Lahore, Punjab, Pakistan

Abstract

We explore the feasibility of reusing fact-checking claims across organizations and languages to improve efficiency and reduce costs. We use an approximate caching approach using multilingual embeddings in a vector database to identify recurring claims. Our method assesses claim recurrence, cache hit rates, and optimal similarity thresholds, enhancing the scalability of global fact-checking processes.

CCS Concepts

• Human-centered computing; **• Collaborative and social computing**; **• Collaborative and social computing systems and tools**; **• Social tagging systems**;

Keywords

Misinformation, social computing, caching, fact checking, embeddings

ACM Reference Format:
Abdullah Ghani, Danish Athar, and Muhammad Ayain Fida Rana. 2024. Embedding The Truth: Approximate Caching for Fact Checking. In *Proceedings of Topics in Large Language Models (LLMs '24)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid spread of misinformation in the digital age poses significant challenges to society, democracy, and public trust [10]. Fact-checking organizations work tirelessly to debunk false claims, but their efforts are hindered by the overwhelming volume of information being output in this day and age. Misinformation does not respect language barriers or regional boundaries—a claim debunked in one country may reappear in another, adapted to different cultural or linguistic contexts [1]. The rapid spread of false claims across social media, news websites, and messaging apps generates a flood of content that must be processed in multiple languages [4]. The repetitive nature of misinformation—where similar claims re-emerge with variations in syntax, phrasing, or framing—further complicates this challenge.

Without automated tools to detect and connect recurring claims efficiently, fact-checkers face duplicated efforts, slower response times, and increased operational costs [6]. These inefficiencies delay the dissemination of accurate information, allowing misinformation to spread unchecked [9]. Mechanisms to identify and reuse previous verifications across languages and regions are crucial for streamlining the fact-checking process and reducing the burden on organizations [3].

Our research attempts to address these challenges by proposing an approximate caching approach using multilingual embeddings and vector databases to identify recurring claims. Our method assesses claim recurrence, cache hit rates, and optimal similarity thresholds, enhancing the scalability of global fact-checking processes. By efficiently reusing verified claims, fact-checkers can reduce the time and resources needed for verification, ensuring that accurate information reaches the public more quickly and mitigating the impact of false narratives.

Furthermore, handling large datasets of claims presents technical challenges. Traditional exact search methods for claim verification reuse become infeasible when dealing with hundreds of thousands of claims [9]. Our use of approximate similarity search using vector databases enables efficient similarity searches, allowing fact-checkers to quickly identify previously verified claims without exhaustive computations. This scalability is essential for managing the growing influx of misinformation. We make our code and dataset publicly available on GitHub [2].

2 Methodology

Our approach identifies recurring claims across languages by leveraging multilingual embeddings and approximate vector similarity search. We fine-tune search parameters, including the k-value, cluster size, and n-probe, to balance precision and efficiency. Finally, we use human verification to evaluate the performance of our proposed scheme. The details. The details of our methodology are given below.

2.1 Multilingual Embeddings

In our approach, we utilized vector embeddings to transform textual data into high-dimensional numerical representations that capture the semantic meaning of the text. This method provides a more accurate way of identifying similar claims compared to traditional keyword matching, which only focuses on exact word matches. By encoding words and sentences based on their contextual relationships, embeddings enable the identification of similar claims even when they differ in wording, syntax, or grammar. This capability ensures that claims expressing the same idea, regardless of their phrasing, are represented by embeddings that are close to

[*] All authors contributed equally to this research.

LLMs '24, Fall 2024, Lahore, PK

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

each other in the vector space, making them ideal for retrieving semantically similar claims.

Given our goal to identify similar claims across different languages, using multilingual embedding models was a natural choice. We decided to use OpenAI's *text-embedding-3-large* since it is designed to generate embeddings that capture nuanced semantic meaning and supports many languages. With its high-dimensional space (8192 dimensions), the model is able to distinguish subtle differences in meaning, ensuring that even claims with similar yet distinct contexts are appropriately differentiated. This approach allowed us to perform similarity searches across claims in multiple languages, capturing their semantic relationships more effectively.

2.2 Vector Database Implementation

We used FAISS (Facebook AI Similarity Search) [5] as our vector database. FAISS is designed for fast and scalable similarity searches on dense vectors, making it ideal for handling large datasets. We used the *flatIVF* (Inverted File) indexing method in FAISS, which balances search speed and memory efficiency by dividing the dataset into clusters and searching within relevant clusters.

2.3 Similarity Search

We used *top-k similarity search* on our vector embeddings in FAISS. Top-k similarity search refers to retrieving the k nearest neighbors for a given query in a vector database. We use it to identify the most similar claims based on their vector embedding representations.

Selecting a k-value: We wanted to balance the efficiency and accuracy of our search, ensuring that relevant claims are retrieved without overwhelming computational resources. A low k-value, such as k = 1, may lead to missed results due to the approximate nature of the search. Conversely, setting k too high increases the computational overhead and may include irrelevant claims, although we apply a minimum similarity threshold for the returned search results to rule out any irrelevant results.

We experimented with various k-values and found that $k = 10$ offers the best balance. This value improves the likelihood of retrieving relevant claims while maintaining search efficiency. Beyond k = 10, search becomes less scalable and does not offer much value either.

2.4 Clustering Strategy:

To organize the embeddings for fast similarity search, we used FAISS to cluster them into approximately $4 \times \sqrt{n}$ clusters, where $n = 130,000$. This approach follows Facebook Research's recommendation of clustering embeddings into $c \times \sqrt{n}$ clusters, where c is a constant factor. Clustering helps reduce the search space, improving efficiency without sacrificing accuracy.

Optimization of n-Probe Parameter The n-probe parameter in FAISS determines how many clusters are searched during a query, impacting both speed and recall. Our literature review indicated that researchers typically use n-probe values between 30 and 50. To refine this further, we systematically iterated over n-probe values ranging from 1 to 100 and tracked the time required to run top-k similarity searches (with $k = 10$) across the entire database. Our findings revealed that even with the largest n-probe value of 100, the time overhead was minimal. Consequently, we selected an n-probe value of 100 for our similarity searches to maximize accuracy while maintaining efficient search performance.

2.5 Thresholding

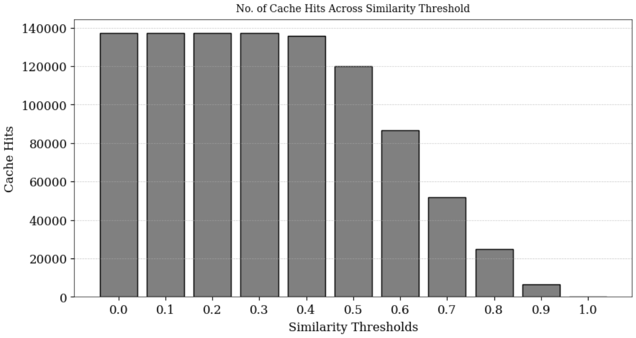

Figure 1: Cache hits for varying minimum similarity thresholds applied to the returned top-k results (k=10). We define a cache hit to be a retrieved claim that meets or exceeds the specified similarity threshold, indicating that it is similar to the input claim.

To ensure that our top-k similarity search results are relevant, we applied a post-retrieval filtering step using a similarity threshold. While the top-k search retrieves the k nearest neighbors (with k = 10), not all results may be semantically relevant due to the approximate nature of similarity searches in large datasets.

Kumar et al.[7] demonstrated that an effective similarity threshold for vector embeddings typically lies between 0.6 and 0.8. An effective threshold is one that strikes a balance between precision (retrieving only relevant claims) and recall (retrieving as many relevant claims as possible). Specifically:

- Thresholds below 0.6 may lead to the inclusion of irrelevant claims (false positives), reducing precision.
- Thresholds above 0.8 may excessively filter out relevant claims (false negatives), reducing recall.

To determine the optimal threshold for our use case, we experimented with various values in the range as shown in 1. The figure illustrates the number of cache hits (retrieved relevant claims) for thresholds ranging from 0.0 to 1.0. Our analysis revealed a clear trade-off between precision and recall as the threshold increased.

Based on these experiments, we selected a threshold of 0.8 to minimize false positives while ensuring that the retrieved claims maintain high semantic relevance.

Manual Verification: To further validate our selection of 0.8 as the minimum similarity threshold, we conducted a manual verification process to assess the relevance of retrieved claims. We selected the top five languages in our dataset based on the number of claims. For each language, we randomly sampled 100 claims and the corresponding top-k search results that had similarity scores exceeding 0.8.

Human raters were tasked with verifying whether the claims returned by the search were indeed similar to the input claims. To ensure the reliability of the evaluation, we measured inter-rater agreement, which reflects the consistency of judgments between raters. The human raters agreed with each other 96.4% of the time, indicating high consistency in their assessments. This high inter-rater agreement is crucial because it validates the objectivity and reliability of the manual verification process.

Additionally, the manual verification revealed that 96.1% of the search results with similarity scores above 0.8 were judged to be similar to the original claim by the human raters, demonstrating that the threshold of 0.8 effectively filters out irrelevant claims while preserving semantically similar ones.

3 Analysis

Building upon the established threshold of 0.8 for our similarity metric, we extended the evaluation across the full dataset of 24,691 claims. Each claim with a nearest-neighbor match above this similarity threshold was considered to have a valid cache hit. Our analysis sought to verify the alignment of fact-checking outcomes between the queried claims and their retrieved semantic matches, investigate the nature of rare mismatches, and explore the impact of language distribution on the performance of our approach.

3.1 Quantitative Evaluation

For each of the 24,691 claims that yielded a cache hit with a similarity score exceeding 0.8, we compared the final fact-checking verdict of the retrieved claim with the verdict associated with the query claim. This step was designed to validate whether retrieving previously verified claims—those deemed semantically similar to the new input—consistently preserves the accuracy of the underlying fact-check. The results revealed a remarkably high concordance: in 99.83% of cases, the verdict of the cache hit (the previously fact-checked claim) matched the ground-truth label of the queried claim. This finding underscores the reliability of leveraging high-similarity claims as proxies to streamline verification tasks. By reducing duplication of effort, fact-checkers can confidently reuse existing evaluations for semantically similar claims, thereby improving efficiency and response times in combating misinformation.

3.2 Qualitative Analysis of Failures

Although the quantitative results are strong, a small subset of claims illustrates the need for caution. These exceptions primarily fall into two categories:

Minor Numerical Deviations: Some retrieved claims closely match the semantics of the queried claim yet contain slight numerical differences. For example:

```
Claim 1: Ghana registered 8.9% GDP growth rate in
    the second quarter of 2021
Claim 2: Ghana recorded a growth rate of 3.1% in
    the first quarter of 2021
```

Although the claims are contextually similar—both pertain to Ghana's GDP growth and share temporal proximity—the numerical discrepancy significantly alters the factual nature of the claim. This can lead to incorrect conclusions if such claims are accepted without further scrutiny.

Temporal Shifts Affecting Claim Veracity: Another source of error arises when two claims describe the same event but at different points in time, thereby changing the claim's truthfulness. Consider the following pair:

```
Claim 1: "Lata Mangeshkar passes away" reviewed on
     19th Nov 2019 (marked False)
Claim 2: "Singer Lata Mangeshkar has passed away"
    reviewed on 22nd Jan 2024 (marked True)
```

While the semantic content is similar, the truth value shifted over time due to real-world events. This highlights a fundamental limitation: claims that are contextually aligned but separated by time cannot always be treated as equivalent.

These findings suggest that while semantic similarity is a powerful tool for automating claim verification, careful temporal and numerical checks are necessary. As a potential mitigation, restricting the acceptable time window for cache hits could help ensure that semantically similar claims align not only in meaning, but also in their temporal context and factual accuracy.

3.3 Language Analysis

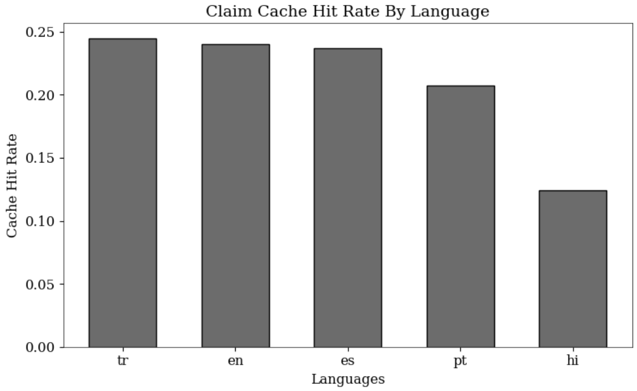

Figure 2: Cache hit rate by language.

We also examined the distribution of cache hits across different languages, as shown in 2. The dataset included claims in Turkish, English, Spanish, and Hindi, among others. Our analysis indicated that Turkish, English, and Spanish exhibited comparable cache hit rates, signaling that the multilingual embeddings perform consistently across these diverse linguistic landscapes. However, the cache hit rate was notably lower for Hindi. We hypothesize that this discrepancy stems from the hyper-localized nature of many Hindi claims, which may reference region-specific entities, cultural nuances, or local events that do not recur frequently in other languages or contexts. The relative scarcity of semantically identical claims across linguistic boundaries may thus reduce the probability of cache hits for these localized claims.

Overall, while the multilingual approach works well for widely spoken languages with more globally shared narratives, it may require additional adaptation or specialized models to handle languages whose fact-checked claims are heavily localized and less likely to have direct semantic counterparts in other linguistic corpora. We believe that this is an important consideration for global fact-checking platforms aiming to scale their approach across highly localized informational ecosystems.

3.4 Temporal Analysis

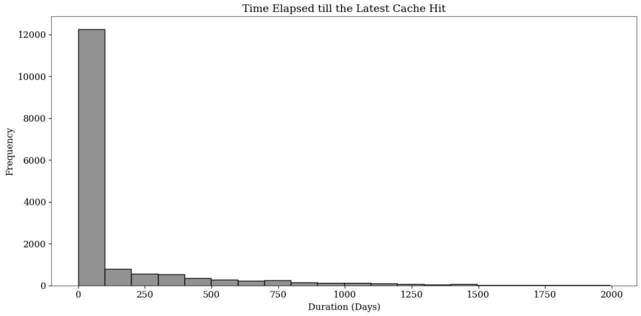

Figure 3: Histogram depicting the distribution of time intervals (in days) between the claim queried and its cache hit claim at a threshold of 0.8. Most cache hits occur shortly before verification, suggesting that many claims are reused within a very narrow time frame.

To further understand the temporal dimension of claim reuse, we examined the distribution of time intervals between a claim's fact-checking and its subsequent cache hits. The histogram in 3 how quickly similar claims tend to reappear and how often previously verified information remains relevant within specific time windows.

3.5 Concentration of Cache Hits in the Short Term

A substantial proportion of cache hits are concentrated within a very short time frame (approaching zero days). This pattern indicates that a large fraction of claims are closely followed by semantically similar claims in the immediate aftermath of their initial verification. Consequently, the rapid recurrence of similar claims supports the utility of similarity-based approaches to fact-checking. By reusing cached results for claims that recur in quick succession, fact-checkers can significantly reduce computational overhead and respond to misinformation more efficiently.

3.6 Long-Tail Occurrences Over Extended Periods

Although most cache hits occur shortly after the initial claim verification, there is a noticeable long tail in the distribution, extending into periods of hundreds or even thousands of days. This indicates that some claims re-emerge long after their initial verification, albeit infrequently. In these cases, the factual landscape may have evolved, and previously determined veracities may no longer hold true.

3.7 Implications for Fact-Checking Practices

For claims that recur after extended intervals, a simple reuse of cached verification results may not be sufficient. Changes in social, political, or technological contexts could alter the claim's veracity. Facts that were once false might become true, or vice versa, depending on real-world developments. To ensure accuracy and relevance in such scenarios, an automated system should initiate a re-verification workflow rather than relying solely on historical cached outcomes.

By striking a balance between rapid cache-based verification for frequently recurring claims and selective re-verification for claims that resurface after prolonged intervals, fact-checking organizations can maintain both efficiency and accuracy. This adaptive approach ensures that updated and contextually relevant information is always available to counter evolving misinformation narratives.

4 Limitations and Future Work

While our approach shows promise, it is not without constraints. One key limitation lies in the handling of temporally sensitive claims. As our analysis revealed, facts can change over time, rendering static cached verifications obsolete. Claims that re-emerge after significant temporal gaps or undergo subtle but critical contextual shifts must be re-verified rather than relying solely on historically cached results. Addressing this challenge may involve integrating temporal context directly into embeddings or dynamically updating cached fact-checks to reflect evolving knowledge bases.

Additionally, our results highlight variability in performance across languages. Although widely spoken languages like English, Turkish, and Spanish displayed comparable cache hit rates, heavily localized languages such as Hindi presented lower reuse potential. This suggests that the generality of multilingual embeddings may be limited for region-specific claims. Future research could explore adaptive or localized embedding models capable of capturing culturally nuanced and niche content more effectively.

Other avenues for improvement include refining similarity thresholds through advanced optimization techniques, incorporating domain-specific ontologies to better discern subtle semantic differences, and experimenting with more robust vector database indexing methods. Beyond technical enhancements, a user-centric approach—such as integrating fact-checker feedback loops or developing interactive interfaces—could further boost system accuracy and utility.

Therefore, while the proposed system provides a scalable and efficient framework for multilingual claim reuse, it remains essential to refine temporal sensitivity, adapt to localized information ecosystems, and integrate continuous improvements as misinformation tactics and linguistic landscapes evolve. The ongoing development of such systems promises to offer more responsive, context-aware, and globally applicable fact-checking solutions.

5 Related Work

A growing body of research has explored the use of semantic similarity and word embeddings to enhance automated fact-checking. ClaimCheck [8] leverages semantic similarity systems to detect repeated false claims by comparing new claims against previously verified ones, thereby improving efficiency during high-stakes events like electoral debates. Our work advances this stream by placing a specific emphasis on multilingual embeddings and the strategic reuse of verified claims, rather than experimenting with a broad spectrum of models or solely focusing on monolingual setups. Whereas prior studies often concentrate on English-language datasets and employ varied model architectures, our approach refines a single embedding-based pipeline to achieve efficient, scalable, and language-agnostic fact-checking. By systematically analyzing and optimizing parameters such as similarity thresholds, we offer a more granular understanding of how multilingual embeddings can streamline and enhance fact-checking workflows across culturally and linguistically diverse information ecosystems.

6 Conclusion

In this work, we presented a multilingual, similarity-based system designed to efficiently identify and reuse verified claims across diverse linguistic and cultural contexts. By leveraging high-dimensional multilingual embeddings and approximate nearest-neighbor search techniques through FAISS, we demonstrated the ability to scale fact-checking operations to large datasets containing tens of thousands of claims. Our approach successfully automated the detection of semantically similar claims, reducing redundant verification efforts and accelerating the dissemination of accurate information. Through careful tuning of similarity thresholds and search parameters, we achieved a high degree of alignment between retrieved cache hits and their corresponding ground-truth verifications, with over 99% agreement in many cases. These findings underscore the potential of embedding-based methodologies to bolster fact-checking workflows, support the ongoing battle against misinformation, and maintain public trust in reliable information sources.

7 Acknowledgments

We thank Dr. Zafar Ayyub Qazi and Dr. Ihsan Ayyub Qazi for introducing us to the process of formal research and introducing us to the various tools we used such as FAISS. We also thank Muhammad Danyal Malik from CSaLT for providing us with the hardware resources for running our experiments.

References

[1] Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, et al. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* 7, 9 (2023), 1502–1513.

[2] Danish Athar. 2024. Topics in LLMs Project. https://github.com/daaanish/TopicsInLLMs-Project Accessed: 2024-12-19.

[3] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* 6, 8 (2024), 852–863.

[4] Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (Jan 2019). https://doi.org/10.1038/s41467-018-07761-2

[5] Dimitrios Danopoulos, Christoforos Kachris, and Dimitrios J. Soudris. 2019. Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud. In *International Conference / Workshop on Embedded Computer Systems: Architectures, Modeling and Simulation*. https://api.semanticscholar.org/CorpusID:199501646

[6] Nicholas Dias and Amy Sippitt. 2020. Researching fact checking: Present limitations and future opportunities. *The Political Quarterly* 91, 3 (2020), 605–613.

[7] Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. 2024. Specious Sites: Tracking the Spread and Sway of Spurious News Stories at Scale . In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1609–1627. https://doi.org/10.1109/SP54263.2024.00171

[8] Irene Larraz, Rubén Míguez, and Francesca Sallicati. 2023. Semantic similarity models for automated fact-checking: ClaimCheck as a claim matching tool. *Profesional de la información* 32, 3 (2023), e320321. https://doi.org/10.3145/epi.2023.may.21

[9] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. https://arxiv.org/abs/1708.07104

[10] Zack Stanton. 2020. You're Living in the Golden Age of Conspiracy Theories. https://www.politico.com/news/magazine/2020/06/17/conspiracy-theories-pandemic-trump-2020-election-coronavirus-326530