# MultiClaimNet: A Massively Multilingual Dataset of Fact-Checked Claim Clusters

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

**Rubén Míguez**
Newtral Media Audiovisual
Spain
`ruben.miguez@newtral.es`

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

## Abstract

In the context of fact-checking, claims are often repeated across various platforms and in different languages, which can benefit from a process that reduces this redundancy. While retrieving previously fact-checked claims has been investigated as a solution, the growing number of unverified claims and expanding size of fact-checked databases calls for alternative, more efficient solutions. A promising solution is to group claims that discuss the same underlying facts into clusters to improve claim retrieval and validation. However, research on claim clustering is hindered by the lack of suitable datasets. To bridge this gap, we introduce *MultiClaimNet*, a collection of three multilingual claim cluster datasets containing claims in 86 languages across diverse topics. Claim clusters are formed automatically from claim-matching pairs with limited manual intervention. We leverage two existing claim-matching datasets to form the smaller datasets within *MultiClaimNet*. To build the larger dataset, we propose and validate an approach involving retrieval of approximate nearest neighbors to form candidate claim pairs and an automated annotation of claim similarity using large language models. This larger dataset contains 85.3K fact-checked claims written in 78 languages. We further conduct extensive experiments using various clustering techniques and sentence embedding models to establish baseline performance. Our datasets and findings provide a strong foundation for scalable claim clustering, contributing to efficient fact-checking pipelines.

## 1 Introduction

Automated fact-checking has become a crucial task to tackle the vast amount of unverified information circulating online. The core objectives of fact-checking pipelines are to identify claims that require verification, retrieve evidence, and assess their veracity automatically. The process can become even more challenging when the same claims are posted on different platforms in different languages (Smeros et al., 2021; Quelle et al., 2023). To overcome this challenge, the claim retrieval component in a fact-checking pipeline can retrieve, for each unverified claim, a set of previously fact-checked claim matches from a database, where available (Panchendrarajan and Zubiaga, 2024).

As the number of verified and unverified claims grows, performing pairwise checks of each new claim against each of the database entries becomes inefficient and impractical for scalable fact-checking pipelines. Since claims are often repeated, an alternative solution is to form claim clusters by grouping the verified/unverified claims discussing the same underlying facts. This not only reduces redundancy in claim retrieval and validation but also enhances the efficiency and scalability of the fact-checking process.

Research on identifying claim clusters has received limited attention in the literature, primarily due to the lack of suitable datasets. Existing studies have applied various clustering techniques to manually verify the existence of claim clusters (Kazemi et al., 2021; Nielsen and McConville, 2022; Quelle et al., 2023). However, to the best of our knowledge no prior work has assessed the quality of the clusters due to the unavailability of datasets. Meanwhile, recent studies (Kazemi et al., 2021; Larraz et al., 2023) have focused on annotating claim pairs that discuss the same underlying facts, enabling a more granular analysis of relationships between two claims. However, extending this to manual annotation of groups of claims that discuss the same fact is a more challenging and time-consuming task, which has hindered creation of datasets.

In this research, we address this challenge by introducing and validating a methodology for data collection and labeling, and by automatically constructing multilingual claim cluster datasets from claim-matching pairs with limited manual intervention. To the best of our knowledge, this is the

first work to create dedicated datasets for claim clustering. We present *MultiClaimNet*, a collection of three multilingual claim cluster datasets. The smaller datasets within *MultiClaimNet* are derived from two existing claim-matching datasets, while the larger dataset is automatically constructed from the fact-checked claim dataset MultiClaim (Pikuliak et al., 2023). Our approach for building this largest dataset involves retrieval of approximate nearest neighbors (ANN) to form candidate claim pairs, followed by automated similarity annotation using three large language models (LLMs). *MultiClaimNet* comprises claims written in 86 unique languages across its three datasets, with the largest dataset containing 85.3K fact-checked claims. Furthermore, we conduct extensive experiments on the three datasets using various clustering approaches in combination with sentence embedding models, including LLMs, to establish baseline performance. We make the following key contributions:

- We present *MultiClaimNet*[1], a collection of three multilingual claim cluster datasets, constructed from claim-matching pairs with minimal manual intervention.
- We automatically generate the largest dataset in *MultiClaimNet* by leveraging a novel data collection and labeling methodology involving ANN retrieval and LLM annotation with no human intervention.
- We conduct extensive experiments with various clustering approaches and sentence embedding models to provide initial insights into the baseline performance.

We believe our datasets and findings will pave the way for further research in claim clustering, contributing to scalable automated fact-checking.

## 2 Related Work

Claim clustering has been less widely studied than other subtasks within automated fact-checking, primarily due to the lack of available datasets. Most existing studies have only explored claim clustering as a means to validate the task and to check if clusters can be found.

Kazemi et al. (2021) applied a single-link hierarchical clustering, a variation of agglomeration clustering algorithm in multilingual claims to confirm the existence of claim clusters. While their study identified meaningful multilingual clusters,

it did not assess the quality of the clusters identified. Hale et al. (2024) followed the same approach to perform an extensive manual analysis of claim clusters found in social media posts related to the Brazilian general election. To address limitations arising from the unavailability of datasets, Adler and Boscaini-Gilroy (2019) assumed that claims related to the same article belonged to the same cluster. They employed the DBSCAN clustering algorithm along with a community detection algorithm (Ester et al., 1996) to form claim clusters. The authors performed a quantitative evaluation by measuring the fraction of claims that belong to the same news article.

The most relevant study, conducted by Quelle et al. (2023), aimed to track claim evolution across languages. The authors represented claims as vectors using sentence embeddings and identified connected components by retrieving the most similar nearest neighbors. These connected components were considered claim clusters. However, none of the works created datasets with ground-truth labels, primarily due to the challenges associated with manual annotation.

Instead of grouping claims discussing the same facts, several studies attempted to analyze the claims discussing the same topic. Nielsen and McConville (2022) employed HDBSCAN (McInnes et al., 2017) on sentence embeddings to obtain more granular clusters discussing the same topic. Smeros et al. (2021) used topic modeling techniques such as LDA (Jelodar et al., 2019) to generate topic vectors and then applied the KMeans algorithm to form the topic clusters. In contrast to these approaches, Shliselberg et al. (2024) leveraged large language models to annotate the topic associated with a claim and then trained a topic classifier in the synthetic dataset. while the topic labels can be used to form topic clusters, the number of distinct topics remains limited.

## 3 Building the Claim Cluster Datasets

We construct the claim cluster datasets from claim matching datasets, which consist of annotated claim pairs labeled as either similar or dissimilar based on whether they discuss the same underlying fact (Panchendrarajan and Zubiaga, 2024). We adhere to the definition of claim similarity provided by Larraz et al. (2023). If two claims are labeled as similar, we assume the similarity is bidirectional; therefore, they belong to the same cluster. This as-

---

| Dataset | # Claim Pairs | # Clusters | # Claims | Avg. Cluster Size | Max Cluster Size | # Language |
|---------|---------------|------------|----------|-------------------|------------------|------------|
| ClaimCheck | 5.2K | 197 | 1187 | 6.03 | 28 | 22 |
| ClaimMatch | 1.5K | 192 | 1171 | 6.1 | 35 | 36 |
| MultiClaim | 54.4K | 30.9K | 85.3K | 2.76 | 54 | 78 |

Table 1: Statistics of *MultiClaimNet*

sumption enables the creation of links between similar claim pairs, forming clusters of interconnected claims. For example, if Claim *A* is annotated as similar to Claim *B*, and Claim *B* is similarly linked to Claim *C*, we infer that Claims *A*, *B*, and *C* belong to the same cluster. We utilized two existing multilingual claim matching datasets along with automatically annotated claim pairs using large language models to form three different claim cluster datasets.

### 3.1 Using Existing Claim Matching Datasets

#### 3.1.1 Data Source

We leveraged the following two claim matching datasets as the sources for creating the first two claim cluster datasets.

*ClaimCheck* (Larraz et al., 2023) - This dataset consists of 7.7K claim pairs annotated for their similarity. The annotations were generated through a combination of manual and automated methods, with a 46-54 ratio. Automated annotation was performed using a similarity-based approach. The dataset primarily includes claims in Spanish and English, along with 20 other languages. For cluster creation, we utilized the 5.2K claim pairs labeled as similar. The remaining pairs labeled as dissimilar were discarded as they are not suitable for forming links between claims.

*ClaimMatch* - This dataset was obtained from the authors of *ClaimCheck* (Larraz et al., 2023) from Newtral Media Audiovisual, a fact-checking organization from Spain. This dataset is relatively smaller, with 2K claim pairs manually annotated. The authors used a set of queries or topics to curate these claims from the Google FactCheck API. For cluster creation, we utilized 1.5K similar pairs written in 36 languages.

#### 3.1.2 Cluster Creation

As mentioned earlier, we automatically generate clusters by linking claims shared among similar claim pairs. Using this approach, we clustered the claims from the two datasets discussed in the previous section, resulting in two multilingual claim clusters. However, this approach has the following two key limitations.

- Missing links - If two similar claims do not appear as similar pairs in the claim-matching dataset, this may not create the link between them, resulting in subclusters discussing the same fact.
- Merging of wrong clusters - If a statement discusses multiple claims, then creating links using such statements may result in the merging of claim clusters discussing different facts.

To avoid the first limitation of missing links, we performed a manual inspection of clusters with high similarity. We converted the English translation of the claim to sentence embedding using Sentence Transformer (Reimers and Gurevych, 2019) and then computed the cluster embedding by averaging the sentence embedding of all claims within a cluster. *ClaimMatch* dataset includes both original and translated claims. For the *ClaimCheck* dataset, we translated the claims into English using Microsoft Azure AI translator (Junczys-Dowmunt, 2019). We then retrieved all the claim clusters with a cosine similarity greater than 0.75 and manually merged them into a single cluster if they discussed the same fact. This process resulted in only a small number of manual merges for both datasets.

To address the second limitation, we manually validated clusters containing more than 20 claims for potential mismerges. However, we found that none of the clusters contained claims discussing different facts across both datasets, which further helped validate our approach. Table 1 presents the statistics of the two datasets. Despite having a higher number of claim pairs in *ClaimCheck*, both datasets exhibit similar data patterns, except for the maximum cluster size and the number of languages present, both of which are slightly higher in *ClaimMatch*.

Figure 1 presents the 2D visualization of the *ClaimCheck* clusters obtained using UMAP dimensionality reduction on the sentence embeddings of claims. Claims within the same cluster are assigned the same color within a subspace (colors reused due to larger number of clusters). Some clusters in the figure are labeled with their respective topics. Interestingly, similar topics or concepts
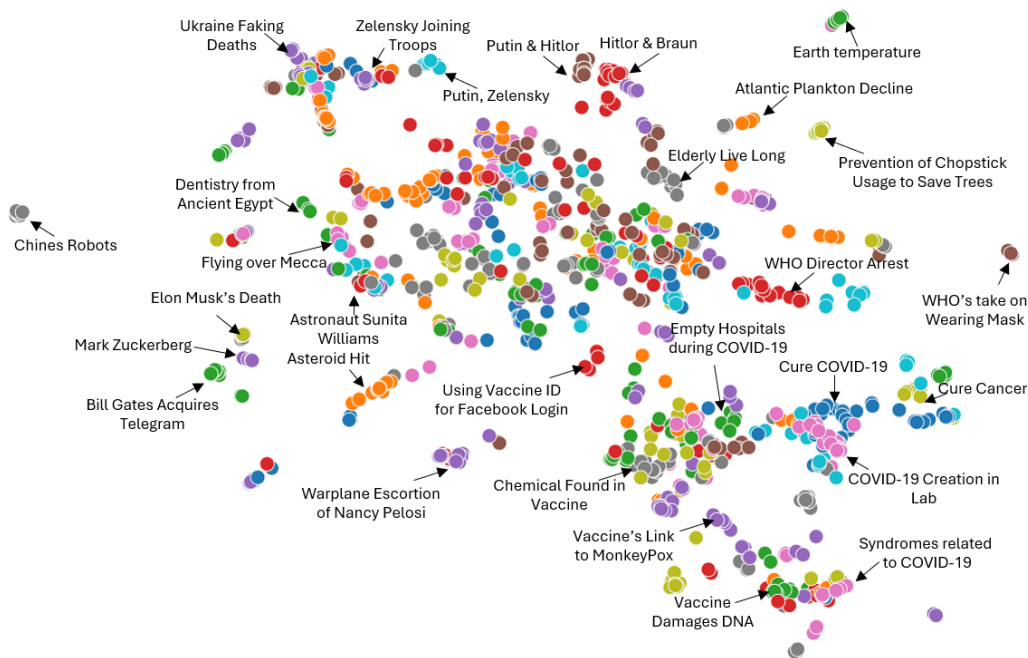
Figure 1: 2D Visualization of ClaimCheck Clusters

are placed closer together in the 2D space, and there exist smooth transitions in topics across the space. For instance, the progression from *Vaccine → COVID-19 → Wearing Mask → WHO → Environmental Issues* illustrates how clusters are interrelated. This suggests that beyond distinct claim clusters, broader conceptual groupings emerge, linking related topics across the space.

## 3.2 MultiClaim Dataset

The two datasets introduced earlier exhibit certain biases due to their curation process. Notably, *ClaimCheck* contains similar pairs with higher semantic similarity, while *ClaimMatch* covers only a limited set of topics. Further, both datasets are relatively small in size, which does not represent a real-world fact-checked database. To address these limitations, we automatically constructed a large-scale claim cluster dataset, mitigating biases related to similarity, topic coverage, and dataset size.

For this study, we utilized the fact-checked multilingual claims from the MultiClaim (Pikuliak et al., 2023) dataset, which comprises 204K claims in 97 languages. The claims were primarily sourced from Google FactCheck Explorer. Given the dataset's extensive size, we propose a two-step approach for automatically constructing claim clusters. Figure 2 illustrates this approach, which is further detailed in the following sections.

### 3.2.1 Sub-Cluster Creation

To identify groups of claims that discuss the same facts, we first need to detect similar claim pairs. Similar claims are likely to be positioned close together in a semantic vector space. Therefore, we employed an Approximate Nearest Neighbor (ANN) search to find the closest claim for each claim in the dataset. Specifically, we use Hierarchical Navigable Small Worlds (HNSW) (Malkov and Yashunin, 2018), one of the most widely used ANN algorithms. Each claim is represented as an embedding vector of its English translation, generated using a Sentence Transformer. We then retrieve the nearest neighbor of each claim as a potential candidate for similar claim pairs. This resulted in 162K unique claim pairs (some claim pairs were duplicated because they were mutually the nearest neighbors of each other) to be annotated as similar or dissimilar. We retrieved only one nearest neighbor, as selecting more would significantly increase the volume of candidate pairs.

Although we retrieved the closest neighbor for each claim, the similarity of the resulting candidate pairs ranged from 0.3 to 1, with an average of 0.71 (Refer to Figure 5 in Appendix for similarity distribution). Among the 162K claim pairs, some of them were an exact translation of the other claim, which we automatically annotated as similar. The remaining 160K claim pairs were annotated as similar or dissimilar using seven large language models
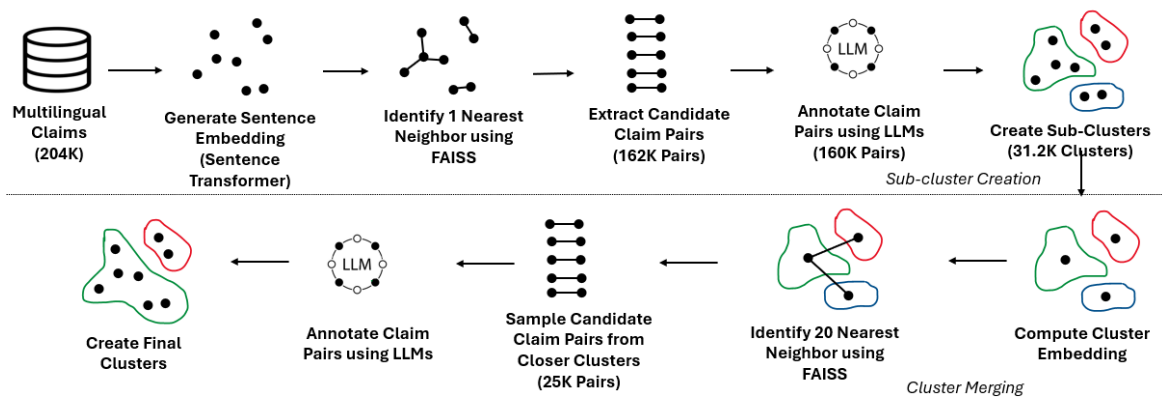
Figure 2: Methodology used for MultiClaim Dataset

(LLMs): Falcon 11B, Falcon 40B, GPT-4, Llama3 8B, Mistral 7B, Mixtral 8×7B, and Phi3 14B. We compared models of varying sizes within the same family to assess their consistency and agreement. We prompted these models with the question:

*Do 'Claim 1' and 'Claim 2' discuss the same claim? Respond with Yes or No.*

We deliberately used the term 'same' instead of 'similar' in the prompt, as the latter introduced more noise. Further, the prompt was not tailored to an LLM to have a generalized and simple prompt formulation that can be applied uniformly across models to identify similar claim pairs. Among the seven LLMs, GPT-4, Phi-3 14B, and Falcon 40B exhibited a higher agreement rate of 77–80% (Refer to Figure 6 in Appendix for agreement rate between LLMs). To minimize noise in cluster creation, we selected only these three models with higher agreement as LLM annotators.

To reduce the impact of potential inaccuracies or hallucinations, we rely on a consensus-based approach: only the claim pairs that were unanimously labeled as similar by all three selected LLMs were retained to ensure label reliability. While this may exclude some positive samples misclassified by one of the models, we found that relaxing the labeling criteria introduced more noise in subsequent steps. This process yielded 54K claim pairs consistently labeled as similar by all three LLMs. We followed a similar approach to construct claim clusters by linking these similar claim pairs. However, since we initially retrieved only one nearest neighbor per claim, this may lead to the splitting of clusters. Therefore, we refer to the generated clusters as sub-clusters and apply a merging process to refine them. In total, 31.2K sub-clusters were constructed at this stage.

### 3.2.2 Cluster Merging

We merge the sub-clusters generated in the previous step to eliminate biases, including the limitation of retrieving only a single nearest neighbor and the strict selection of claim pairs annotated as similar by all three LLMs. First, we represent each sub-cluster as an embedding by averaging the sentence embeddings of its claims. Next, we retrieve the top 20 nearest neighbors of each sub-cluster using HNSW and sample one claim per cluster to form candidate claim pairs. Since our clustering criteria is based on links between claim pairs, we believe that retrieving 20 nearest neighbors is sufficient to identify missing links between sub-clusters. Similar to the manual validation, we filtered out claim pairs from clusters with a cosine similarity of less than 0.75. This process yields 25K claim pairs, which are then for their similarity by LLMs.

During the annotation of inter-cluster claim pairs, we observed a decrease in label agreement across LLMs, ranging from 41% to 69%. This suggests that annotating inter-cluster claim pairs was more challenging for the LLMs compared to the annotation of nearest neighbors. However, we only considered the 8.5K claim pairs annotated with the same label by all three LLMs as valid, as relaxing this criterion led to mismerges. This process yielded a final set of 30.9K claim clusters. We refer to this dataset with its original name, *MultiClaim*, and its statistics are detailed in Table 1.

### 3.3 Multilingual Claims

Our *MultiClaimNet* dataset encompasses multilingual claims written in 86 unique languages across the three cluster datasets. Refer to Appendix A.3 for the language-wise statistics of the datasets. The

| Dataset | Monolingual vs Multilingual Clusters | Avg. Unique Languages in Multilingual Clusters |
|---|---|---|
| ClaimCheck | 55/142 | 3.2 |
| ClaimMatch | 58/134 | 3.8 |
| MultiClaim | 15.9K/15K | 2.4 |

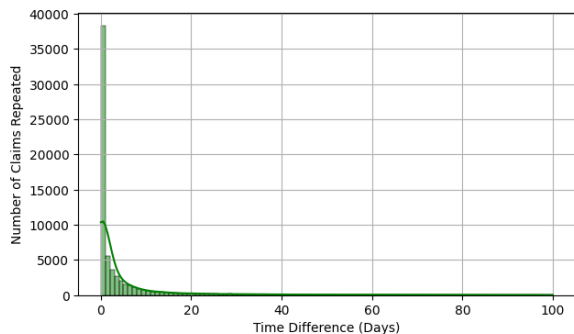Table 2: Statistics of Multilingual Clusters



Figure 3: Number of Claims Repeated over First 100 Days in *MultiClaim* Clusters

smaller datasets exhibit similar language distributions, with Spanish being the dominant language in the *ClaimCheck* dataset, and English as the majority language in *ClaimMatch*. Additionally, both of these datasets predominantly feature European languages. In contrast, *MultiClaim* includes a broader mix of both Asian and European languages, with English remaining the dominant language.

Table 2 presents the statistics of multilingual clusters present in the cluster datasets. Notably, the smaller datasets contain a higher proportion of multilingual clusters, whereas the larger one has an equal number of monolingual and multilingual clusters. Furthermore, the multilingual claim clusters consist of claims written in 2.4-3.8 unique languages on average, validating the existence of similar claims across different languages. Further, this highlights the importance of advancing multilingual research in automated fact-checking to effectively handle repeated claims across diverse linguistic contexts.

### 3.4 Temporal Claims

The reoccurrence of the same claim demands techniques to tackle repeated claims for an effective fact-checking pipeline. Among the three datasets, *MultiClaim* contains the timestamps of the claims, making it an ideal choice for further research on temporal claims. Figure 3 illustrates the distribution of claim repetitions over the first 100 days. Notably, 50% of the claims are repeated within just

1.6 days, and 75% are repeated within the first 29.3 days from the occurrence of the first claim in a cluster. This underscores the need for advanced methods to detect and manage repeated claims efficiently, ensuring that fact-checking systems can respond swiftly and accurately to the rapid spread of misinformation.

## 4 Experiment Setup

We employ various clustering techniques to evaluate the performance of baseline models on the claim cluster datasets.

### 4.1 Clustering Approaches

In real-world fact-checked databases, the number of clusters is often unknown. Therefore, we utilize clustering techniques that do not require the number of clusters as a predefined parameter. Instead, these methods automatically determine the optimal number of clusters based on other controlling parameters, such as density thresholds or distance metrics. We apply the following clustering methods:

- HDBSCAN (McInnes et al., 2017) - A hierarchical density-based clustering algorithm that determines dense regions and merges them to form hierarchical trees.
- Agglomerative clustering (Müllner, 2011) - A hierarchical clustering method that builds tree-like structures and iteratively merges them using links.
- Affinity Propagation (Dueck, 2009) - A message-passing clustering algorithm that identifies representative points and generates clusters by assigning other points to them based on similarity.
- Birch (Zhang et al., 1996) - A scalable algorithm designed for large datasets to form tree structures and merge trees iteratively.
- MeanShift (Comaniciu and Meer, 2002) - A centroid-based clustering algorithm that iteratively shifts points toward high-density regions to form clusters automatically
- Optics (Ankerst et al., 1999) - A density-based clustering algorithm that identifies dense regions from points ordered according to their reachability distance.

The hyperparameters used for the clustering algorithms and detailed in Appendix A.4. Density-based approaches are shown to be very effective when combined with dimensionality reduction tech-

11208

| Model | Parameters | Embedding |
|---|---|---|
| Distiluse-base-multilingual-cased | 135M | 512 |
| Paraphrase-multilingual-MiniLM-L12 | 118M | 768 |
| Paraphrase-multilingual-mpnet-base | 278M | 768 |
| Gte-multilingual-base | 305M | 768 |
| All-roberta-large-v1 | 355M | 1024 |
| LaBSE | 471M | 768 |
| KaLM-embedding-multilingual-mini-instruct-v1.5 | 494M | 896 |
| Multilingual-e5-large-instruct | 560M | 1024 |
| Bge-m3 | 567M | 1024 |
| Gte-Qwen2-1.5B-instruct | 1B | 1536 |
| MiniCPM-Embedding | 2.4B | 2304 |
| E5-mistral-7b-instruct | 7B | 4096 |
| Gte-Qwen2-7B-instruct | 7B | 3584 |
| LLM2Vec-Llama-3-8B-Instruct-mntp | 8B | 4096 |
| Bge-multilingual-gemma2 | 9B | 3584 |

Table 3: Multilingual Sentence Embedding Models

niques such as UMAP (Nielsen and McConville, 2022). Therefore, we reduce the sentence embedding to 8 dimensions (optimal for all datasets) for HDBSCAN and Optics.

## 4.2 Multilingual Sentence Embedding Models

Unlike dataset curation, where we used the English translation, we explore multilingual representations of claims by encoding their original text as sentence embeddings. In our experiments, we evaluate fifteen multilingual sentence embedding models listed in Table 3.

## 4.3 Metrics

We report the following metrics, which measure different aspects of the clusters generated against the ground-truth clusters (Pauletic et al., 2019).

- Adjusted Rand Index (ARI) - Measures similarity between two clusters
- Adjusted Mutual Index (AMI) - Measures mutual information shared between two clusters
- Homogeneity (HMG) - Measures the fraction of cluster instances belonging to the same ground-truth cluster
- Completeness (CMP) - Measures the fraction of ground-truth cluster instances that are grouped together
- V-Measure (VM) - Measures the harmonic mean of homogeneity and completeness
- Purity - Measures chances of cluster instances belonging to the same ground truth cluster

## 5 Results

### 5.1 Clustering Approaches

Table 4 presents the performance of different clustering approaches across the three datasets. For this experiment, we use the *Bge-m3* (Chen et al., 2024) model, as it performed well across all datasets. Results for the Affinity propagation and MeanShift approaches are not reported for *MultiClaim* due to high memory and running time requirements. However, Affinity propagation generally performs well across the smaller dataset, achieving the highest scores across multiple metrics.

Agglomerative clustering is the only approach that consistently performed well in the largest dataset across all metrics. This highlights its potential for integration in large-scale fact-check databases. However, its performance slightly declines on smaller datasets as they tend to generate more fine-grained clusters (i.e., a larger number of clusters than the ground truth), leading to higher homogeneity and purity scores. HDBSCAN efficiently clusters both small and large datasets. However, in *MultiClaim*, it tends to perform poorly, as the reduction in dimensionality consistently leads to excessively granular clusters.

Although *ClaimMatch* contains a similar number of claims and clusters, it shows significant drops in ARI and AMI, suggesting substantial overlap between clusters. In contrast, *ClaimCheck* exhibits clearer semantic separation, possibly due to highly similar claim pairs, allowing the baselines to easily achieve 99% homogeneity and 93% completeness. These results highlight that while the baselines can effectively handle *ClaimCheck*, they struggle to resolve the ambiguity present in *ClaimMatch*, despite its smaller size, and face even greater challenges with the large and complex *MultiClaim* dataset.

### 5.2 Multilingual Sentence Embedding Models

Table 5 presents the performance of various sentence embedding models when applied with Agglomerative clustering. Interestingly, smaller models (< 1B parameters) achieved the highest scores compared to the larger models. In particular, *gte-multilingual-base* (Zhang et al., 2024) achieves the highest performance across small datasets, while *Bge-m3* (Chen et al., 2024) outperforms others on the *MultiClaim* dataset. This suggests that large language models are not necessarily required for effective sentence representation in semantic similarity tasks such as clustering.

However, we observed that clustering algorithms are highly sensitive to sentence representation unless combined with a dimensional reduction technique. For instance, Table 7 presents the average number of clusters and their corresponding stan-

| Dataset | Approach | # Clusters | ARI | AMI | HMG | CMP | V-Measure | Purity |
|---|---|---|---|---|---|---|---|---|
| ClaimCheck | HDBScan | 188 | 0.794 | 0.89 | 0.952 | 0.96 | 0.956 | 0.893 |
| | Agglomerative | 275 | 0.723 | 0.874 | **0.994** | 0.922 | 0.956 | **0.987** |
| | AffinityPropagation | 183 | **0.806** | **0.9** | 0.956 | **0.963** | **0.96** | 0.919 |
| | Birch | 196 | 0.551 | 0.827 | 0.894 | 0.957 | 0.925 | 0.796 |
| | MeanShift | 212 | 0.116 | 0.629 | 0.743 | 0.916 | 0.821 | 0.637 |
| | Optics | 258 | 0.051 | 0.53 | 0.749 | 0.857 | 0.799 | 0.702 |
| ClaimMatch | HDBScan | 191 | 0.584 | 0.779 | 0.908 | 0.913 | 0.91 | 0.819 |
| | Agglomerative | 292 | 0.591 | 0.791 | **0.978** | 0.885 | **0.93** | **0.946** |
| | AffinityPropagation | 158 | **0.68** | **0.819** | 0.915 | **0.934** | 0.924 | 0.833 |
| | Birch | 99 | 0.269 | 0.64 | 0.688 | 0.931 | 0.792 | 0.437 |
| | MeanShift | 186 | 0.041 | 0.413 | 0.588 | 0.83 | 0.688 | 0.439 |
| | Optics | 284 | 0.11 | 0.537 | 0.828 | 0.821 | 0.825 | 0.759 |
| MultiClaim | HDBScan | 10599 | 0.007 | 0.383 | 0.733 | **0.975** | 0.837 | 0.339 |
| | Agglomerative | 27853 | **0.574** | **0.714** | **0.961** | 0.973 | **0.967** | **0.784** |
| | Birch | 32034 | 0.071 | 0.398 | 0.873 | 0.946 | 0.908 | 0.605 |
| | Optics | 21509 | 0.001 | 0.221 | 0.71 | 0.935 | 0.807 | 0.519 |

Table 4: Performance of Different Clustering Approaches

| | ClaimCheck | | | ClaimMatch | | | MultiClaim | | |
|---|---|---|---|---|---|---|---|---|---|
| | ARI | AMI | VM | ARI | AMI | VM | ARI | AMI | VM |
| Distiluse-base-multilingual-cased | 0.544 | 0.752 | 0.924 | 0.425 | 0.635 | 0.894 | 0.222 | 0.477 | 0.939 |
| Paraphrase-multilingual-MiniLM-L12 | 0.657 | 0.821 | 0.941 | 0.552 | 0.743 | 0.915 | 0.331 | 0.564 | 0.947 |
| Paraphrase-multilingual-mpnet-base | 0.702 | 0.854 | 0.95 | 0.606 | 0.788 | 0.927 | 0.488 | 0.639 | 0.957 |
| Gte-multilingual-base | **0.845** | **0.919** | **0.969** | **0.711** | **0.842** | **0.941** | 0.467 | 0.656 | 0.952 |
| All-roberta-large-v1 | 0.411 | 0.634 | 0.892 | 0.328 | 0.522 | 0.855 | 0.111 | 0.344 | 0.908 |
| LaBSE | 0.586 | 0.772 | 0.928 | 0.462 | 0.672 | 0.899 | 0.506 | 0.631 | 0.96 |
| KaLM-embedding-multilingual-mini-instruct-v1.5 | 0.703 | 0.826 | 0.925 | 0.489 | 0.686 | 0.86 | 0.138 | 0.371 | 0.872 |
| Multilingual-e5-large-instruct | 0.732 | 0.85 | 0.931 | 0.545 | 0.738 | 0.875 | 0.132 | 0.391 | 0.867 |
| Bge-m3 | 0.723 | 0.874 | 0.956 | 0.591 | 0.791 | 0.93 | **0.574** | **0.714** | **0.967** |
| Gte-Qwen2-1.5B-instruct | 0.587 | 0.796 | 0.935 | 0.478 | 0.708 | 0.909 | 0.443 | 0.569 | 0.952 |
| MiniCPM-Embedding | 0.585 | 0.783 | 0.931 | 0.439 | 0.654 | 0.893 | 0.301 | 0.489 | 0.938 |
| Gte-Qwen2-7B-instruct | 0.665 | 0.835 | 0.945 | 0.51 | 0.732 | 0.915 | 0.493 | 0.634 | 0.958 |
| E5-mistral-7b-instruct | 0.821 | 0.905 | 0.963 | 0.626 | 0.789 | 0.921 | 0.268 | 0.465 | 0.914 |
| LLM2Vec-Llama-3-8B-Instruct-mntp | 0.285 | 0.47 | 0.786 | 0.208 | 0.372 | 0.744 | 0.045 | 0.219 | 0.836 |
| Bge-multilingual-gemma2 | 0.607 | 0.798 | 0.933 | 0.507 | 0.73 | 0.913 | 0.438 | 0.591 | 0.952 |

Table 5: Performance of Different Multilingual Sentence Embedding Models

dard deviation across different sentence embedding models combined with HDBSCAN and Agglomerative clustering. Agglomerative exhibits greater sensitivity to the sentence representation, leading to a 4-7.5 fold increase in standard deviation compared to HDBSCAN. This suggests that developing more robust clustering solutions, independent of the number of data instances and sentence representations, is essential for future research.

## 5.3 Error Analysis

Table 6 lists some of the error scenarios found across all three datasets during the error analysis of the clusters formed using baselines. Except for the last, *merge due to common entities*, other scenarios result in misplacing claims discussing the same facts in different clusters. This highlights the limitations of the baseline approaches and suggests the need for task-specific solutions for obtaining accurate claim clusters.
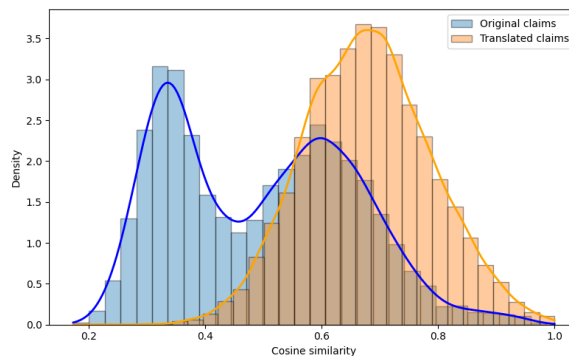


Figure 4: Similarity Distribution of Misplaced Claims and their Closest Counterparts from Baseline Clusters in *MultiClaim*. Blue: Original Claims; Orange: English Translation.

Beyond the above misgrouped cases, we observed that the baseline struggles to cluster highly similar claims across languages in the *MultiClaim* dataset, possibly due to the linguistic diversity

| Scenario | Misgrouped Claims (English Translation) |
|---|---|
| Not recognizing the same entities with different reference | AstraZeneca vaccine against covid-19 causes monkeypox<br>Monkeypox is a consequence of anti-covid vaccination |
| Not knowing the background of the entities | Tedros Adhanom was arrested by Interpol<br>WHO Chief Arrested For Crimes Against Humanity |
| Not focusing on keywords | The American Rescue Plan helped create nearly 10 million new jobs.<br>Joe Biden states that the American Rescue Plan helped create nearly 10 million new jobs. |
| Prone to noise | A satirical article virally shared via social media 'reported' that World Health Organization director Tedros Adhanom Ghebreyesus had been arrested for crimes against humanity.<br>Tedros Adhanom was arrested by Interpol |
| Not recognizing implicit claim | Picture shows Zelenskyy's Russian passport.<br>Zelensky has a Russian passport and not a Ukrainian passport |
| Merge due to common entities | Maduro has threatened to send missiles to Spain<br>Maduro asks for support for Brazil's military to overthrow Bolsonaro |

Table 6: Error Scenarios and Sample Misgrouped Claims

| | HDBSACN | Agglomerative |
|---|---|---|
| ClaimCheck | 207 ± 19 | 276 ± 92 |
| ClaimMatch | 215 ± 14 | 290 ± 104 |
| MultiClaim | 8.5K ± 2.3K | 22.1K ± 9.4K |

Table 7: Average Number of Clusters Found

present. Figure 4 illustrates the similarity distribution of misplaced claims with respect to their closest counterparts within the baseline clusters. We compute cosine similarity using both the original and the translated claims. While the majority of original claims exhibit lower similarity scores (average of $0.5 \pm 0.16$), their corresponding English translations yield much higher similarity (average of $0.68 \pm 0.1$). These differences in distribution indicate that the model has difficulty identifying semantically similar claims across languages, underscoring the need for stronger multilingual models.

## 6 Applications

Although claim clustering can be integrated across various components of automated fact-checking, we outline key applications of our datasets.

- Claim cluster database - Verified claims can be maintained as claim clusters by grouping the claims discussing the same facts. Traditional clustering techniques can be further explored to group multilingual claims at a larger scale.
- Fact-checked claim cluster retrieval - Instead of retrieving individual verified claims from a fact-checked dataset, an unverified claim can be matched against an entire cluster, offering a broader context of the claim across different occurrences of the claim in different languages.
- Iterative claim clustering - Clustering approaches can be explored further to perform iterative clustering representing real-world sce-

narios of integration of new claims, where it can be merged into an existing cluster or form a new cluster.
- Verification of claim clusters - Similar to the grouping of verified claims, unverified claims can be clustered together, streamlining the verification process by allowing entire clusters to be assessed collectively.
- Visualize topic themes - Claim clusters can be further grouped using techniques such as hierarchical clustering to analyze inter-related claim clusters and to identify topic themes.

## 7 Conclusion

This paper introduces *MultiClaimNet*, a collection of three multilingual claim clustering datasets constructed from similar claim pairs. The largest dataset within *MultiClaimNet* was created by retrieving claim pairs using approximate nearest neighbor approaches and annotating them with large language models. We conduct extensive experiments on these datasets, evaluating different clustering approaches and multilingual sentence embedding models. Our results show that Agglomeration clustering performs well on the largest dataset and remains competitive across the smaller ones. In contrast, HDBSCAN excels only on the smaller datasets, underscoring the need for robust solutions. Additionally, the smaller sentence embedding models outperform larger ones, highlighting their potential for scalable fact-checking solutions. Our error analysis further suggests that task-specific approaches are essential for improving clustering accuracy. We further highlight the potential applications of our dataset and believe it will be a valuable resource for advancing multilingual fact-checking research.

## Limitations

The limitations of this work are as follows:

- Data curation bias - As discussed earlier, the smaller dataset introduces biases such as highly similar claim pairs and a limited range of topics due to the data curation techniques.
- Reliance on textual content alone for annotation - Claim pairs in *MultiClaim* dataset were annotated for their similarity based solely on textual content presented to LLMs. This may result in false positives when other media associated with a claim refer to different incidents.
- Mismerge due to text with multiple claims - We assume that a fact-checked claim generally discusses one factual statement. However, the presence of claims with more than one factual statement may create false positive links between claim pairs.

## Acknowledgments

## References

Ben Adler and Giacomo Boscaini-Gilroy. 2019. Real-time claim detection from news articles and retrieval of semantically-similar factchecks. *arXiv preprint arXiv:1907.02030*.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

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.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Delbert Dueck. 2009. *Affinity propagation: clustering data by passing messages*. Ph.D. thesis.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Scott A Hale, Adriano Belisario, Ahmed Nasser Mostafa, and Chico Camargo. 2024. Analyzing misinformation claims during the 2022 brazilian general election on whatsapp, twitter, and kwai. *International Journal of Public Opinion Research*, 36(3):edae032.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, et al. 2025. Kalmembedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A. Hale. 2021. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

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

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).

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*.

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.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

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.

Iva Pauletic, Lucia Nacinovic Prskalo, and Marija Brkic Bakaric. 2019. An overview of clustering models with an application to document clustering. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1659–1664. IEEE.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Mária Bieliková. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500.

Dorian Quelle, Calvin Cheng, Alexandre Bovet, and Scott A Hale. 2023. Lost in translation–multilingual misinformation and its evolution. *arXiv preprint arXiv:2310.18089*.

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.

Michael Shliselberg, Ashkan Kazemi, Scott A Hale, and Shiri Dori-Hacohen. 2024. Syndy: Syn thetic dy namic dataset generation framework for misinformation tasks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2801–2805.

Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2021. Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1692–1702.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

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

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

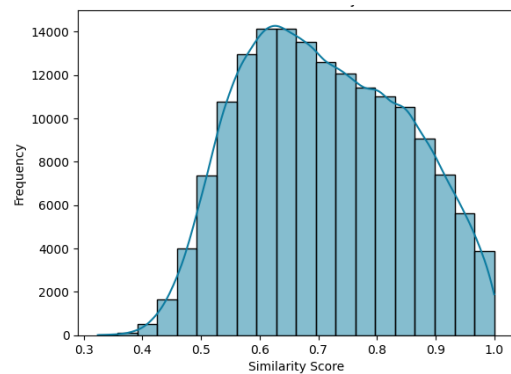# A Appendix

## A.1 Statistics



Figure 5: Similarity Distribution of Claim Pairs Extraction for Sub-cluster Creation

Figure 5 illustrates the similarity distribution of the nearest neighbors retrieved during the sub-clustering stage as described in 3.2.1. The similarity of nearest neighbors ranges from 0.31 to 1.

Figure 6 presents the percentage of same label annotated by the seven LLMs. Among the seven LLMs compared, Falcon20B, GPT4, and Phi3-14B exhibited a higher percentage of agreement ranging from 77.38% to 80.63.

## A.2 Sample Claims

Table 8 presents sample claims from the largest cluster in each dataset. Claims about the same fact vary in length and tend to refer to the same

| Dataset | Factchecked Claim (English Translation) |
|---------|----------------------------------------|
| ClaimCheck | COVID-19 vaccines contain graphene oxide |
| | Former Pfizer employee confirms vaccine contains graphene oxide nanoparticles |
| | A Facebook post claimed that Spanish doctors had detected graphene oxide in Covid vaccines |
| ClaimMatch | This video proves chemtrail |
| | Harmful substances are being dropped from airplanes. |
| | The claim that the video shows the moment when the pilot mistakenly released harmful chemicals referred to as chemtrails at the airport |
| MultiClaim | Video shows Russian forces attacking Ukraine |
| | A video of a recent Russian airstrike on Ukraine |
| | The video shows a large explosion that took place in Ukraine after Russian President Vladimir Putin announced a military operation in Ukraine |

Table 8: Sample FactChecked Claims from the Largest Clusters
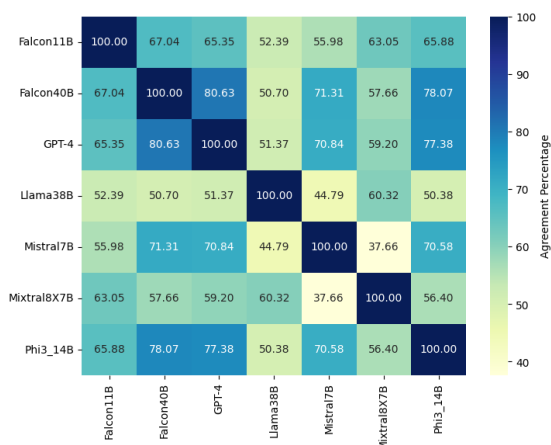


Figure 6: Percentage of Same Label Annotation among LLMs during Sub-cluster Creation

entities related to the claim with various references. As mentioned in Limitations (Section 7), the LLM-based annotation for claim similarity may introduce false positive samples when the media associated with the claims refer to different incidents. For example, in *MultiClaim*, the largest cluster consists of claims about Russian attacks in Ukraine. While these claims share similar textual descriptions, they may refer to different events within the war.

## A.3 Language Statistics

Figure 7 presents the language statistics across three datasets, including only languages that appear at least three times in a dataset.

## A.4 Hyperparameter Tuning

We use the Scikit-learn[2] library for the execution of clustering algorithms and the UMAP library[3] for dimension reduction. The optimal hyperparameters that yield the best performance in terms of V-Measure across all three datasets were used to re-

[2]https://scikit-learn.org/stable/api/sklearn.cluster.html
[3]https://umap-learn.readthedocs.io/

port the results. Table A.4 lists the hyperparameter search and the optimal settings used. Apart from these hyperparameters, we use the default parameters recommended in the library for each clustering approach.

## A.5 Implementation

Baseline performances were obtained using 1 GPU (Volta V100 or Ampere A100) with 8 CPU cores, each composed of 11 GB memory. We used the following multilingual sentence embedding models available in HuggingFace for the analysis of baseline performance.

- Distiluse-base-multilingual-cased[4] (Reimers and Gurevych, 2019)
- Paraphrase-multilingual-MiniLM-L12[5] (Reimers and Gurevych, 2019)
- Paraphrase-multilingual-mpnet-base[6] (Reimers and Gurevych, 2019)
- Gte-multilingual-base[7] (Zhang et al., 2024)
- All-roberta-large-v1[8] (Reimers and Gurevych, 2019)
- LaBSE[9] (Reimers and Gurevych, 2019)
- KaLM-embedding-multilingual-mini-instruct-v1.5[10] (Hu et al., 2025)
- Multilingual-e5-large-instruct[11] (Wang et al., 2024b)
- Bge-m3[12] (Chen et al., 2024) (Chen et al., 2024)

[4]https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1
[5]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[6]sentence-transformers/paraphrase-multilingual-mpnet-base-v2
[7]Alibaba-NLP/gte-multilingual-base
[8]sentence-transformers/all-roberta-large-v1
[9]sentence-transformers/LaBSE
[10]HIT-TMG/KaLM-embedding-multilingual-mini-\instruct-v1.5
[11]intfloat/multilingual-e5-large-instruct
[12]https://huggingface.co/BAAI/bge-m3
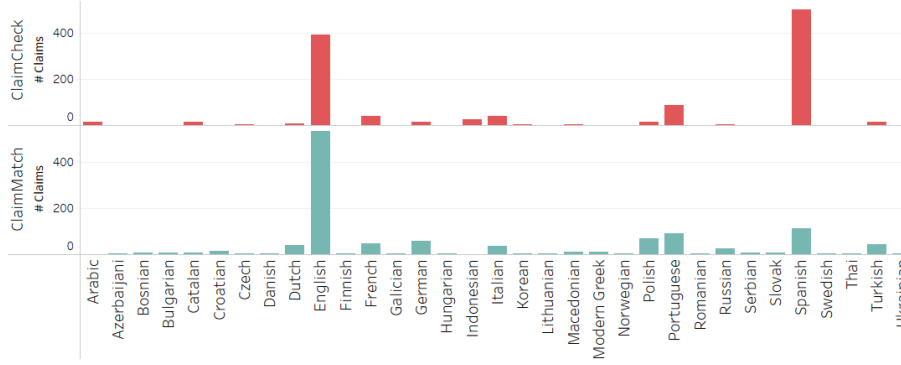
(a) ClaimCheck and ClaimMatch
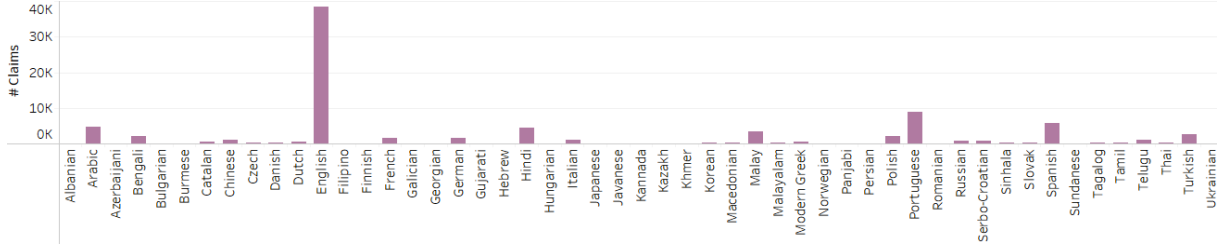


(b) MultiClaim

Figure 7: Statistics of Language Occurrence in the *MultiClaimNet* Dataset

| Approach | Hyperparameter | Search Options | Optimal |
|---|---|---|---|
| HDBSAN | vector dimenstion | [2,3,4,8,16,32,64,128,256,512] | 8 |
| | cluster selection epsilon | Minimum value | 0.1 |
| | min samples | Minimum value | 1 |
| | min cluster size | Minimum value | 2 |
| Agglomerative | distance threshold | [0.5 - 2], step size of 0.5 | 1 |
| | linkage | [ward, complete, average, single] | ward |
| Birch | threshold | [0.1 - 1], step size 0.1 | 0.7 |
| Meanshift | bandwidth | [0.1 - 0.8], step size of 0.05 | 0.75 |
| OPTICS | min samples | Minimum value | 2 |

Table 9: Hyperparameters of Clustering Approaches

- Gte-Qwen2-1.5B-instruct[13] (Li et al., 2023)
- MiniCPM-Embedding[14] (Hu et al.)
- E5-mistral-7b-instruct[15] (Wang et al., 2024a)
- Gte-Qwen2-7B-instruct[16] (Li et al., 2023)
- LLM2Vec-Llama-3-8B-Instruct-mntp[17] (BehnamGhader et al., 2024)
- Bge-multilingual-gemma2[18] (Chen et al., 2024)

We used all-MiniLM-L6-v2[19] model (Reimers and Gurevych, 2019) to convert the English translations of the claim to sentence embedding during the data curation process (Section 3).

| | ClaimCheck | ClaimMatch | MultiClaim |
|---|---|---|---|
| HDBScan | 0.001 | 0.001 | 0.388 |
| Agglomerative | 0.005 | 0.006 | 42.276 |
| AffinityPropagation | 0.007 | 0.017 | - |
| Birch | 0.001 | 0.001 | 0.329 |
| MeanShift | 0.387 | 0.585 | - |
| Optics | 0.032 | 0.037 | 4.372 |

Table 10: Running Time of the Clustering Algorithms in Minutes

## A.6 Running Time

Table 10 lists the running time of the clustering algorithms across three datasets.