# A Heterogeneous Graph Benchmark for Misinformation on Twitter

Dan S. Nielsen
dan.nielsen@bristol.ac.uk
Department of Engineering Mathematics
University of Bristol
UK

Ryan McConville
ryan.mcconville@bristol.ac.uk
Department of Engineering Mathematics
University of Bristol
UK

## ABSTRACT

Misinformation is becoming increasingly prevalent on social media and in news articles. It has become so widespread that we require algorithmic assistance utilising machine learning to detect such content. Training these machine learning models require datasets of sufficient scale, diversity and quality. However, datasets in the field of automatic misinformation detection are predominantly monolingual, include a limited amount of modalities and are not of sufficient scale and quality. Addressing this, we develop a data collection and linking system (`MuMiN-trawl`), to build a public misinformation graph dataset (`MuMiN`), containing rich social media data (tweets, replies, users, images, articles, hashtags) spanning 21 million tweets belonging to 26 thousand Twitter threads, each of which have been semantically linked to 13 thousand fact-checked claims across dozens of topics, events and domains, in 41 different languages, spanning more than a decade. The dataset is made available as a heterogeneous graph via a Python package (`mumin`). We provide baseline results for two node classification tasks related to the veracity of a claim involving social media, and demonstrate that these are challenging tasks, with the highest macro-average F1-score being 62.55% and 61.45% for the two tasks, respectively. The `MuMiN` ecosystem is available at https://mumin-dataset.github.io/, including the data, documentation, tutorials and leaderboards.

## CCS CONCEPTS

• **Mathematics of computing** → *Graph algorithms*; Cluster analysis; • **Information systems** → *Network data models*; • **Computing methodologies** → *Supervised learning by classification.*

## KEYWORDS

dataset, misinformation, graph, twitter, social network, fake news

**ACM Reference Format:**
Dan S. Nielsen and Ryan McConville. 2022. A Heterogeneous Graph Benchmark for Misinformation on Twitter. In *Proceedings of Workshop on Graph Learning Benchmarks (GLB '22).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*GLB '22, April 26, 2022, Virtual*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

While it may be possible to track the history of misinformation, or 'fake news', back to Octavian of the Roman Republic [34], or Browne in the 17th century [6], it was the World Wide Web and the rise of online social networks that has provided new and powerful ways for the rapid dissemination of information, both true and false, with false information having negative effects across many aspects of society, such as politics and health.

Given the severity of online misinformation, there have been numerous public datasets made available for researchers to develop and evaluate automated misinformation detection models. These publicly available datasets cover topics ranging over celebrities [25], rumours [36], politics [28] and health [19]. These datasets typically include data from a social network, usually Twitter, along with labels assigning them to a category, categorising them as some equivalent of 'true' or 'false'. These labels often come from 'fact-checking' resources such as PolitiFact[1] and Poynter[2].

There are, however, a number of limitations of existing datasets. We believe that in order to make advances on the development of automated misinformation detection systems, datasets that capture the breadth, complexity and scale of the problem are required. Specifically, we believe that an effective dataset should be large scale, as misinformation is an extremely varied and wide ranging phenomenon, with thousands of manually fact-checked claims available online from fact-checking organisations across a range of topics. To ensure that misinformation detection models are able to generalise to new events, we need models to be able to learn event-independent predictors of misinformation. We believe that such predictors will not be possible from the claim texts alone, as they are inherently event-dependent. Instead, we argue that models (and thus datasets to train them) should utilise the context of the claim, for example, the social network surrounding the claim, or the article in which the claim was posted. Further, given that misinformation is a global challenge, a useful dataset should not be limited to a single language, and should contain data in as many languages as possible.

We see the goal of an automatic misinformation detection system as a tool that can help people identity misinformation so that they can act on it accordingly. Considering that a lot of the misinformation today is spread on social media networks, such a system should be able to retrieve, connect and utilise the information in these networks to identify misinformation as accurately as possible. This is the core rationale behind our proposed two tasks, which we further discuss in Section 5.1:

[1] https://www.politifact.com/
[2] https://www.poynter.org/ifcn/

(1) Determine the veracity of a claim, given its social network context.
(2) Determine the likelihood that a social media post to be fact-checked is discussing a misleading claim, given its social network context.

To this end, we present the dataset MuMiN, which addresses the limitations of existing work. In summary, our main contributions are as follows:

- We release a graph dataset, MuMiN, containing rich social media data (tweets, replies, users, images, articles, hashtags) spanning 21 million tweets belonging to 26 thousand Twitter threads, each of which have been semantically linked to 13 thousand fact-checked claims across dozens of topics, events and domains, in 41 different languages, spanning more than a decade.
- We release the data collection and linking system, MuMiN-trawl, which was used to build the MuMiN dataset.
- We release a Python package, mumin, which eases the compilation of the dataset as well as enabling easy export to the Deep Graph Learning framework [32].
- We propose two representative tasks involving claims and social networks. We provide baseline results considering both text-only models, image-only models as well as using a heterogeneous graph neural network.

## 2 DATASET CREATION

The dataset creation consists of two parts, the first one concerning the claims and their fact-checked verdicts, and the second part concerning the collection of the surrounding social context. The general strategy is to collect claims as spatiotemporally diverse as possible, and to collect as many high-quality social features surrounding these as possible. The dataset creation was performed using MuMiN-trawl on a single workstation with an Intel Core i9-9900K CPU, 64GB of RAM, with two Nvidia 2080Ti GPUs, with the collection taking several months. Baseline results were produced on the same workstation.

For the collection of fact-checked claims we utilise the Google Fact Check Tools API[3], which is a resource that collects fact-checked claims from fact-checking organisations around the world. This API was also used in Shiao and Papalexakis [27] to create a dataset for automatic misinformation detection, but our aim was to collect a much larger amount of claims that were sufficiently diverse, both in terms of content and language. We compiled a list of 115 fact-checking organisations and collected all the fact-checked claims for each of them, from the fact-checking organisation's inception up until present day. This resulted in 128,070 fact-checked claims.

The first challenge is that the verdict is unstructured freetext and can be written in any language at any length. To remedy this, we trained a 'verdict classifier', a machine learning model that classifies the freetext verdicts into three pre-specified categories: misinformation, factual and other. Towards this, we manually labelled 2,500 unique verdicts. We trained both a monolingual English model on translated verdicts as well as a multilingual model, with the English model performing marginally better (0.99 validation macro-F1 among the misinformation and factual labels, compared to 0.98 for the multilingual model). As the English-only model was marginally better than the multilingual model, we opted to use that in building the dataset. However, we appreciate the convenience of not having to translate the verdicts, so we release both the English-only and multilingual verdict classifiers on the Hugging Face Hub[4]. See [23] for more details regarding the verdict classifier, and see the appendix for some examples of the verdicts and resulting predicted verdicts. With the performance satisfactory, we then used the model to assign labels to all of the plaintext verdicts in the dataset.

From the claims and verdicts, we next collected relevant social media data. This data was collected from Twitter[5] using their Academic Search API[6], where we aimed to collect as many relevant Twitter threads that shared and discussed content related to the claims obtained through the method described above. We extracted five keyphrases for each claim[7], and queried the Twitter Academic Search API for the first 100 results for each keyphrase, where we required the results to not be replies, had to share either a link or an image, and had to have at least 5 retweets. This resulted in approximately 2.5 million tweets.

From the database of tweets, the next task was to find all the Twitter threads that were relevant to each claim. We translated all the claims, tweets and articles into English and embedded them using the same model. We then computed cosine similarities between the claims and tweets as well as the claims and articles.

The resulting cosine similarity distribution can be found in the appendix. We decided to release three datasets, corresponding to the three thresholds 0.7, 0.75 and 0.8. These thresholds were chosen based on a qualitative evaluation of a subset of the linked claims; see examples of such linked claims at various thresholds in the appendix. The lower threshold dataset is of course larger, but also contains more label noise, whereas the higher threshold dataset is considerably smaller, but with higher quality labels. See various statistics of these datasets in Table 1.

From the resulting Twitter posts linked to the claims we next queried Twitter for the surrounding context of these posts. We retrieved, for each tweet, a sample of 500 replies to the tweet and 500 quote tweets of the tweet (along with their authors), 100 users that retweeted the tweet, 100 users who followed the authors of the tweet, 100 users who were followed by the author of the tweet and all users who was mentioned in the tweet. For each of these users, we further queried Twitter for their recent 100 tweets.

See Nielsen and McConville [23] for a more detailed description of the construction of the MuMiN dataset.

## 3 DATASET DESCRIPTION

Given the scale and diversity of the data collected it is not possible to succinctly provide a thorough analysis, which we leave to future

[3]https://developers.google.com/fact-check/tools/api/reference/rest

[4]See https://hf.co/saattrupdan/verdict-classifier-en and https://hf.co/saattrupdan/verdict-classifier.

[5]https://www.twitter.com

[6]https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

[7]This was done using the KeyBERT [11] package together with the paraphrase-multilingual-MiniLM-L12-v2 model from the Sentence Transformer package [26].

**Table 1: The statistics of the three datasets.**

| Dataset | #Claims | #Threads | #Tweets | #Users | #Articles | #Images | #Languages | %Misinfo |
|---|---|---|---|---|---|---|---|---|
| MuMiN-large | 12,914 | 26,048 | 21,565,018 | 1,986,354 | 10,920 | 6,573 | 41 | 94.79% |
| MuMiN-medium | 5,565 | 10,832 | 12,659,371 | 1,150,259 | 4,212 | 2,510 | 37 | 94.20% |
| MuMiN-small | 2,183 | 4,344 | 7,202,506 | 639,559 | 1,497 | 1,036 | 35 | 92.71% |

work, and other researchers interested in exploring and using our dataset. Nonetheless, we will provide a preliminary analysis of various aspects of the dataset.

As mentioned in Section 2, we release three datasets, corresponding to the cosine similarity thresholds 0.7, 0.75 and 0.8. The statistics of the datasets can be found in Table 1. Note the heavy class imbalance of the datasets, which is likely due to the fact that fact-checking organisations are more interested in novel claims, and these tend to favour misinformation [31]. A common way to fix this issue [7, 19] is to collect news articles from "trusted sources" and use tweets connected to these as a means to increase the `factual` class. However, as these will likely arise from a different distribution than the rest of the datasets (they might not be novel claims, say), we decided against that and left the dataset as-is. We have instead released the source code we used to collect the dataset, `MuMiN-trawl`, which can be used to collect extra data, if needed[8].

To adhere to the terms and conditions of Twitter, the dataset will only contain the tweet IDs and user IDs, from which the tweets and the user data can be collected via the Twitter API using our `mumin` package (see Section 4). Further, to comply with copyright restrictions of the fact-checking websites, we do not release the claims themselves. Instead, we release claim and cluster keyphrases, the former obtained as described in Section 2 and the latter obtained as described in Section 3.1. The datasets thus contain the tweet IDs, user IDs and claim keyphrases, as well as the `POSTED`, `MENTIONS`, `FOLLOWS`, `DISCUSSES` and `IS_REPLY_TO` relations, shown in the appendix. From these, the remaining part of the dataset can be built by using our `mumin` package, see Section 4.

## 3.1 Claim Topic Clusters

We performed clustering on embeddings of the claim text in order to extract higher level topics or events from the claims. Using a `UMAP` [22] projection of embeddings of the claims and `HDBSCAN` [20], a hierarchical density based clustering algorithm, we were able to discover 26 clusters based on the claim text. We optimised the hyperparameters of the projection as well as the clustering algorithm[9], achieving a silhouette coefficient of 0.28. The clusters can be seen in the appendix.

To provide context for each cluster, we concatenated the claims in each cluster and extracted keyphrases from each cluster[10]. From these, it is apparent that the claims can be clustered into diverse topics, ranging from COVID-19 (a cluster of approximately half of all claims), to topics ranging from natural disasters to national and international political and social events. These "cluster keyphrases" have been included for each claim in the dataset.

# 4 THE `MUMIN` PACKAGE

As we can only release the tweet IDs and user IDs to adhere to Twitter's terms of use, we have built a Python package, `mumin`, to enable compilation of the dataset as easily as possible. The package can be installed from PyPI using the command `pip install mumin`, and the dataset can be compiled as follows:

```
>>> from mumin import MuminDataset
>>> dataset = MuminDataset(bearer_token, size='small')
>>> dataset.compile()
```

Here `bearer_token` is the Twitter API bearer token, which can be obtained from the Twitter API website. The `size` argument determines the size of the dataset to load and can be set to 'small', 'medium' or 'large'. Further, there are many arguments included in the `MuminDataset` constructor which controls what data to include in the dataset. For instance, one can set `include_tweet_images` to `False` to not include any images[11].

With the dataset compiled, the graph nodes can be accessed through `dataset.nodes` and the relations can be accessed through `dataset.rels`. A convenience method `dataset.to_dgl` returns a heterogeneous graph object to be used with the DGL library [32].

We have built a tutorial on how to use the compiled dataset, including building different classifiers. We also release the source code for the `mumin` package[12].

# 5 MODEL PERFORMANCE

## 5.1 Baseline Models

The `MuMiN` dataset lends itself to several different classification tasks, relating the various modalities to the verdicts of the associated claims (`misinformation` or `factual`). As mentioned in Section 1, we have chosen to provide baselines related to the following two tasks:

(1) Given a claim and its surrounding subgraph extracted from social media, predict whether the verdict of the claim is `misinformation` or `factual`. We name this task "claim classification".
(2) Given a source tweet (i.e., not a reply, quote tweet or retweet) to be fact-checked, predict whether the tweet discusses a claim whose verdict is `misinformation` or `factual`. We name this task "tweet classification".

[8] This can be found at https://mumin-dataset.github.io/.
[9] This optimisation resulted in the hyperparameters `n_neighbors=50`, `n_components=100`, `random_state=4242` and `metric='cosine'` for `UMAP`, and `min_samples=15` and `min_cluster_size=40` for `HDBSCAN`. This was done using the Python packages `scikit-learn` [24] and `hdbscan` [21].
[10] This was done using the KeyBERT library [11] on embeddings produced by a Sentence Transformer `paraphrase-multilingual-MiniLM-L12-v2` [26].
[11] See https://mumin-build.readthedocs.io for a full list of arguments.
[12] The tutorial and all the source code can be accessed through https://mumin-dataset.github.io/.

**Table 2: Dataset split statistics**

| Dataset | %Train | %Val | %Test | %MisinfoTrain | %MisinfoVal | %MisinfoTest | #ClustersTrain | #ClustersVal | #ClustersTest |
|---|---|---|---|---|---|---|---|---|---|
| MuMiN-large | 78.52% | 11.39% | 10.09% | 94.37% | 96.73% | 95.92% | 8 | 21 | 8 |
| MuMiN-medium | 76.98% | 11.61% | 11.41% | 93.79% | 96.73% | 94.46% | 7 | 18 | 7 |
| MuMiN-small | 77.90% | 11.35% | 10.75% | 91.82% | 97.15% | 94.42% | 7 | 15 | 6 |

We implement several baseline models to demonstrate the predictive power of the different modalities for these tasks. Firstly, we implement the LaBSE transformer model from Feng et al. [10] with a linear classification head, and apply this model directly to the claims and the source tweets, respectively. Secondly, we implement the vision transformer (ViT) model from Dosovitskiy et al. [8], also with a linear classification head, and apply this to the subset of the tweets that include images (preserving the same train/val/test splits).

As for a graph baseline, we implement a heterogeneous version of the GraphSAGE model from [13], as follows. For each of the nodes in the dataset (see the appendix for the full graph schema), we sample 100 edges of each edge type connected to it (in any direction), process each of the sampled neighbouring nodes through a GraphSAGE layer, and sum the resulting node representations. Finally, layer normalisation [2] is applied to the aggregated node representations. The baseline model contains two of these graph layers. This graph baseline is trained on MuMiN without profile images, article images and timelines (i.e., tweets that users in our graph have posted, which are not directly connected to any claim)[13]. We call this baseline model HeteroGraphSAGE.

To enable consistent benchmarking on the dataset, we provide train-val-test splits of the data. These have been created such that the splits are covering distinct events, identified by the claim clusters in Section 3.1. Statistics for each of the splits can be found in Table 2, which shows that we still roughly maintain the label balance throughout all the dataset splits.

See Table 3 and 4 for an overview of the performance of each of these models. We see that both tasks are really challenging, with the HeteroGraphSAGE model achieving the best performance overall, but with the text-only LaBSE model not far behind. We note that the HeteroGraphSAGE model only makes two "hops" through the graph, meaning that it is not able to capture all the information that is present in the graph. Increasing the number of hops resulted in poorer performance, which is the well-known "oversmoothing" problem [18, 35].

We have created an online leaderboard containing the results of these baselines and invite researchers to submit their own models. We release all the source code we used to conduct the baseline experiments.[14].

**Table 3: Baseline test performance on the claim classification task, measured in macro-average F1-score (larger is better). Best result for each dataset marked in bold.**

| Model | MuMiN-small | MuMiN-medium | MuMiN-large |
|---|---|---|---|
| Random | 40.07% | 38.96% | 38.79% |
| Majority class | 47.56% | 48.06% | 48.13% |
| LaBSE | **62.55%** | 55.85% | 57.90% |
| HeteroGraphSAGE | 57.95% | **57.70%** | **59.80%** |

**Table 4: Baseline test performance on the tweet classification task, measured in macro-average F1-score (larger is better). Best result for each dataset marked in bold. Note that the ViT model is only trained and evaluated on the subset of the tweets containing images.**

| Model | MuMiN-small | MuMiN-medium | MuMiN-large |
|---|---|---|---|
| Random | 37.18% | 37.72% | 36.90% |
| Majority class | 48.77% | 48.56% | 48.87% |
| ViT | 53.20% | 52.00% | 48.70% |
| LaBSE | 54.50% | **57.45%** | 52.80% |
| HeteroGraphSAGE | **56.05%** | 54.10% | **61.45%** |

## 6 CONCLUSION

In this paper we presented MuMiN, a multilingual graph misinformation dataset containing rich social media data spanning 21 million tweets belonging to 26 thousand Twitter threads, each of which have been semantically linked to 13 thousand fact-checked claims across dozens of topics, events and domains, spanning more than a decade. We also presented a data collection and linking system, MuMiN-trawl. The freetext multilingual verdicts were categorised into the consistent categories of factual or misinformation, using a finetuned transformer model which we also release. We further developed a Python package, mumin, which enables simple compilation of MuMiN as well as providing easy export to the Deep Graph Library. Finally, we proposed and provided baseline results for two node classification tasks. The baselines include text-only and image-only approaches, as well as a heterogeneous graph neural network. We showed that the tasks are challenging, with the highest macro-average F1-score being 62.55% and 61.45% for the two tasks, respectively. The data, along with tutorials and a leaderboard, can be found at https://mumin-dataset.github.io/.

## ACKNOWLEDGMENTS

This research is supported by REPHRAIN: The National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online, under UKRI grant: EP/V011189/1.

[13]Note that, as the graph baseline has two layers, leaving these out does not change the claim classification score, only potentially the tweet classification score.

[14]See https://mumin-dataset.github.io/ for both the leaderboard and the baseline repository.

# REFERENCES


[1] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *EMNLP*.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 21–27. https://doi.org/10.18653/v1/N18-2004

[4] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. 2015. Verifying Multimedia Use at MediaEval 2015. In *MediaEval 2015 Workshop*.

[5] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. 2016. Verifying Multimedia Use at MediaEval 2016. In *MediaEval 2016 Workshop*.

[6] Thomas Browne. 1646. Pseudodoxia Epidemica or Enquiries into very many received tenents and commonly presumed truths. London : printed for Edward Dod, and are to be sould by Andrew Crook.

[7] Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User Preference-aware Fake News Detection. *arXiv preprint arXiv:2104.12259* (2021).

[10] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852* (2020).

[11] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[12] Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A New Benchmark Dataset for Multilingual 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 2: Short Papers)*. Association for Computational Linguistics, Online, 675–682. https://doi.org/10.18653/v1/2021.acl-short.86

[13] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 1025–1035.

[14] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 493–503. https://doi.org/10.18653/v1/K19-1046

[15] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[16] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[17] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task Learning for Rumour Verification. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3402–3413.

[18] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*.

[19] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A Multilingual and Multidimensional Data Repository for CombatingCOVID-19 Fake New. *arXiv e-prints* (2020), arXiv–2011.

[20] Leland McInnes and John Healy. 2017. Accelerated Hierarchical Density Based Clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 33–42.

[21] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (2017), 205.

[22] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[23] Dan Saattrup Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. *arXiv preprint arXiv:2202.11684* (2022).

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[25] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3391–3401.

[26] 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. http://arxiv.org/abs/1908.10084

[27] William Shiao and Evangelos E Papalexakis. 2021. KI2TE: Knowledge-Infused InterpreTable Embeddings for COVID-19 Misinformation Detection. *1st International Workshop on Knowledge Graphs for Online Discourse Analysis, KnOD 2021* (2021).

[28] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[29] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709* 8 (2017).

[30] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 809–819.

[31] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559 arXiv:https://science.sciencemag.org/content/359/6380/1146.full.pdf

[32] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J. Smola, and Zheng Zhang. 2019. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. *CoRR* abs/1909.01315 (2019). http://arxiv.org/abs/1909.01315

[33] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 422–426.

[34] Carol A Watson. 2018. Information literacy in a fake/false news world: An overview of the characteristics of fake news and its historical development. *International Journal of Legal Information* 46, 2 (2018), 93–96.

[35] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81.

[36] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International Conference on Social Informatics*. Springer, 109–123.

# A SUPPLEMENTARY TABLES AND FIGURES

**Figure 2: The distribution of cosine similarities among tweet-claim pairs.**

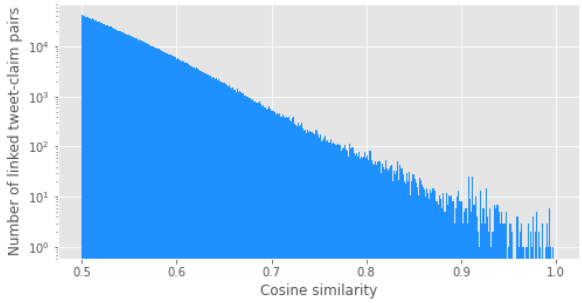


**Figure 3: UMAP projection of the claim text embeddings. The large cluster on the right corresponds to COVID-19 related claims.**

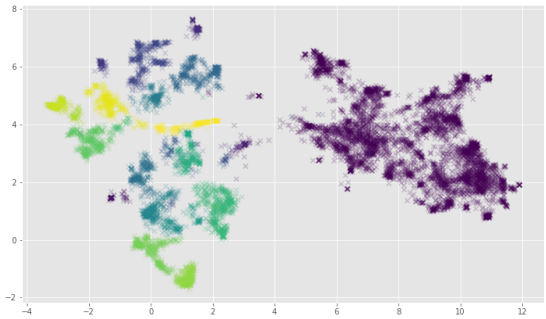

**Table 6: The distribution of the top languages in the MuMiN-large dataset.**

| Language | Proportion | #Claims | %misinfo |
|---|---|---|---|
| English | 42.88% | 5,538 | 92.85% |
| Portuguese | 10.98% | 1,418 | 95.28% |
| Spanish | 8.26% | 1,067 | 95.41% |
| Hindi | 6.16% | 796 | 100.00% |
| Arabic | 4.34% | 560 | 95.18% |
| French | 3.46% | 447 | 97.99% |
| German | 2.91% | 376 | 97.61% |
| Indonesian | 2.55% | 329 | 99.70% |
| Italian | 2.33% | 301 | 89.37% |
| Bengali | 2.26% | 292 | 100.00% |
| Turkish | 2.19% | 283 | 95.41% |
| Polish | 1.73% | 224 | 83.48% |
| Other | 9.93% | 1,283 | 95.49% |

**Table 7: The distribution of the top languages in the MuMiN-medium dataset.**

| Language | Proportion | #Claims | %misinfo |
|---|---|---|---|
| English | 45.46% | 2,530 | 92.29% |
| Portuguese | 10.75% | 598 | 96.49% |
| Spanish | 7.82% | 435 | 94.25% |
| Hindi | 6.50% | 362 | 100.00% |
| Arabic | 4.40% | 245 | 93.88% |
| French | 3.61% | 201 | 97.51% |
| Italian | 3.04% | 169 | 86.98% |
| German | 2.57% | 143 | 97.90% |
| Indonesian | 2.07% | 115 | 100.00% |
| Bengali | 1.99% | 111 | 100.00% |
| Turkish | 1.90% | 106 | 94.34% |
| Polish | 1.40% | 106 | 80.77% |
| Other | 8.48% | 472 | 97.03% |

**Table 8: The distribution of the top languages in the MuMiN-small dataset.**

| Language | Proportion | #Claims | %misinfo |
|---|---|---|---|
| English | 47.41% | 1,035 | 90.34% |
| Portuguese | 10.86% | 237 | 97.47% |
| Spanish | 7.42% | 162 | 92.59% |
| Hindi | 6.92% | 151 | 100.00% |
| Arabic | 4.90% | 107 | 89.72% |
| Italian | 4.49% | 98 | 86.73% |
| French | 3.71% | 81 | 97.53% |
| Turkish | 1.83% | 40 | 87.50% |
| German | 1.51% | 33 | 100.00% |
| Indonesian | 1.51% | 33 | 100.00% |
| Bengali | 1.42% | 31 | 100.00% |
| Polish | 1.15% | 25 | 80.00% |
| Other | 6.87% | 150 | 96.00% |

**Table 9: Sample predictions of the verdict classifier.**

| factual | misinformation | other |
|---|---|---|
| True | False | Satire |
| Correct Attribution | Misleading | Landmarks |
| Broadly correct. | Mostly false | Questionable |
| According to the most recent data, this is about right | Pants on fire | More complex than that |
| This is correct for relative poverty in the UK, measured after housing costs in 2015/16. It's a smaller other measures of poverty. | Three Pinocchios | This video filmed in Equatorial Guinea shows a student attacking one of his teachers |

**Figure 1: The graph schema of the MuMiN dataset.**

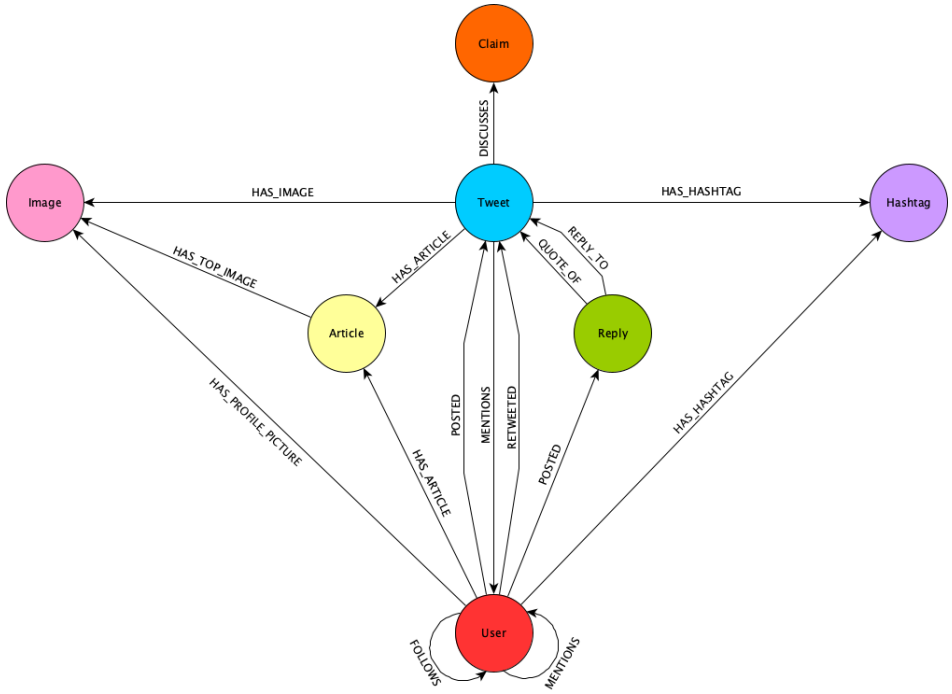


**Table 5: An overview of publicly available datasets for automatic misinformation detection, ordered by release date. Here † indicates that the tweet content is not available but that the related users are, and parentheses indicate that it only holds for a subset of the dataset.**

| Dataset | #Facts | #Tweets | Verified | Multilingual | Multitopical | Articles | Images | User | Social | Replies |
|---|---|---|---|---|---|---|---|---|---|---|
| MediaEval15 [4] | 413 | 15,821 | ✓ | | ✓ | | ✓ | ✓ | | |
| MediaEval16 [5] | 542 | 18,049 | ✓ | | ✓ | | ✓ | ✓ | | |
| Liar [33] | 12,836 | | ✓ | | | | | ✓ | | |
| Weibo [16] | | 9,528 | ✓ | | ✓ | | ✓ | ✓ | | |
| PHEME5 [36] | | 5,802 | ✓ | | ✓ | | | | | ✓ |
| FNN-BuzzFeed [29] | 182 | | ✓ | | | ✓ | | ✓ | ✓ | |
| FNN-PolitiFact17 [29] | 240 | | ✓ | | | ✓ | | ✓ | ✓ | |
| PHEME9 [17] | | 6,425 | ✓ | | ✓ | | | | | ✓ |
| Celebrity [25] | 200 | | ✓ | | | ✓ | | | | |
| FakeNewsAMT [25] | 240 | | | | ✓ | ✓ | | | | |
| FEVER [30] | 185,445 | | | | ✓ | | | | | |
| AFCSDC [3] | 422 | | ✓ | | ✓ | ✓ | | | | |
| UKP Snopes [14] | 6,422 | | ✓ | | ✓ | ✓ | | | | |
| MultiFC [1] | 34,918 | | ✓ | | ✓ | ✓ | | | | |
| HoVer [15] | 26,000 | | | | ✓ | | | | | |
| FNN-PolitiFact20 [28] | 1,056 | 564,129 | ✓ | | | ✓ | ✓ | ✓ | | ✓ |
| FNN-GossipCop [28] | 22,140 | 1,396,548 | ✓ | | | ✓ | ✓ | ✓ | | ✓ |
| CoAID [7] | 4,251 | 160,667 | (✓) | | | ✓ | | | | ✓ |
| MM-COVID [19] | 11,565 | 105,300 | (✓) | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |
| UPFD-POL [9] | 314 | 40,740† | ✓ | | | ✓ | | ✓ | | ✓† |
| UPFD-GOS [9] | 5,464 | 308,798† | ✓ | | | ✓ | | ✓ | | ✓† |
| X-FACT [12] | 31,189 | | ✓ | ✓ | ✓ | | | | | |
| MuMiN | 12,914 | 21,565,018 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 10: Examples of claim-article linking.**

| Translated Claim | Translated Title | Article URL | Similarity |
|---|---|---|---|
| Google removed the term "Palestine" from Google Maps | Google and Apple remove Palestine from their maps | https://bit.ly/mumi-3 | 84.93% |
| China loses control of part of its space rocket, and it will soon fall to Earth. | Heads Up! A Used Chinese Rocket Is Tumbling Back to Earth This Weekend. | https://bit.ly/mumi-4 | 80.47% |
| Photo shows Aung San Suu Kyi being detained during a military coup on February 1, 2021 | Myanmar's army detains Aung San Suu Kyi and government leaders in a possible coup | https://bit.ly/mumi-5 | 75.03% |
| One of the nurses who made the Pfizer-BioNtech vaccine immediately fainted from a side effect of the vaccine. Also, the nurse who fainted after having just been vaccinated is dead. | Live Nurse Faints After Being Vaccinated Against Covid-19! | https://bit.ly/mumi-6 | 70.29% |
| Americans Need WHO COVID-19 Vaccine Card for International Travel | 'Vaccine passport' will define tourism in the world, but countries bar some immunizers | https://bit.ly/mumi-7 | 65.30% |

**Table 11: The 115 fact-checking organisations present in the dataset. The numbers in parentheses indicate how many claims were processed from the website in total.**

| Website | Claims included | Website | Claims included | Website | Claims included |
|---|---|---|---|---|---|
| politifact.com | 716 (7,865) | factcheck.kz | 90 (776) | thip.media | 19 (134) |
| factcheck.afp.com | 581 (4,874) | correctiv.org | 87 (1,313) | scroll.in | 18 (73) |
| boomlive.in | 407 (3,149) | faktograf.hr | 86 (680) | faktisk.no | 17 (640) |
| factual.afp.com | 363 (2,913) | newschecker.in | 83 (1,143) | ici.radio-canada.ca | 17 (102) |
| snopes.com | 361 (4,025) | fatabyyano.net | 77 (1,218) | fakenews.pl | 17 (163) |
| misbar.com | 328 (4,641) | animalpolitico.com | 66 (850) | thejournal.ie | 16 (83) |
| factly.in | 317 (4,113) | factcheck.thedispatch.com | 64 (177) | malayalam.factcrescendo.com | 15 (245) |
| dpa-factchecking.com | 298 (1,474) | lemonde.fr | 62 (564) | factnameh.com | 15 (387) |
| vishvasnews.com | 298 (5,974) | bol.uol.com.br | 62 (407) | factrakers.org | 13 (147) |
| factcheck.org | 268 (2,312) | factcheckthailand.afp.com | 58 (252) | factograph.info | 12 (253) |
| factuel.afp.com | 243 (2,710) | projetocomprova.com.br | 57 (406) | watson.ch | 11 (39) |
| facta.news | 230 (1,196) | noticias.uol.com.br | 56 (693) | poynter.org | 9 (49) |
| fullfact.org | 226 (3,302) | sprawdzam.afp.com | 54 (299) | br.de | 9 (121) |
| thequint.com | 223 (1,084) | dogrulukpayi.com | 53 (641) | mygopen.com | 8 (440) |
| observador.pt | 207 (1,284) | aap.com.au | 52 (365) | factcheckni.org | 8 (141) |
| aajtak.in | 189 (1,539) | newsweek.com | 48 (196) | hindi.asianetnews.com | 8 (165) |
| piaui.folha.uol.com.br | 187 (6,060) | tamil.factcrescendo.com | 47 (1,523) | abc.net.au | 7 (112) |
| newtral.es | 178 (2,353) | periksafakta.afp.com | 47 (415) | liberation.fr | 7 (97) |
| checamos.afp.com | 165 (1,073) | chequeado.com | 46 (1,689) | theconversation.com | 6 (54) |
| polygraph.info | 157 (1,128) | nytimes.com | 44 (497) | telugu.newsmeter.in | 6 (280) |
| aosfatos.org | 155 (1,795) | poligrafo.sapo.pt | 42 (3,496) | factchecker.in | 6 (32) |
| teyit.org | 154 (2,421) | boombd.com | 39 (381) | open.online | 5 (23) |
| usatoday.com | 154 (884) | fakty.afp.com | 38 (220) | bbc.co.uk | 5 (43) |
| politica.estadao.com.br | 151 (1,632) | dailyo.in | 36 (729) | tenykerdes.afp.com | 5 (36) |
| factcrescendo.com | 145 (896) | presseportal.de | 35 (466) | namibiafactcheck.org.na | 4 (36) |
| thelogicalindian.com | 139 (994) | youturn.in | 35 (1,591) | factcheckmyanmar.afp.com | 4 (79) |
| washingtonpost.com | 138 (1,304) | 20minutes.fr | 33 (255) | observers.france24.com | 4 (54) |
| cekfakta.com | 135 (4,104) | altnews.in | 31 (4,996) | oglobo.globo.com | 4 (50) |
| bangla.boomlive.in | 131 (1,640) | cbsnews.com | 30 (231) | buzzfeed.com | 2 (25) |
| ellinikahoaxes.gr | 131 (1,120) | napravoumiru.afp.com | 29 (172) | bangla.aajtak.in | 2 (129) |
| newsmeter.in | 127 (1,430) | semakanfakta.afp.com | 29 (198) | istinomer.rs | 2 (887) |
| boatos.org | 125 (1,893) | faktencheck.afp.com | 27 (335) | verify-sy.com | 2 (56) |
| maldita.es | 123 (1,063) | tjekdet.dk | 27 (481) | thewhistle.globes.co.il | 2 (65) |
| colombiacheck.com | 118 (802) | cinjenice.afp.com | 26 (227) | azattyq.org | 1 (9) |
| demagog.org.pl | 115 (3,181) | vistinomer.mk | 25 (370) | radiofarda.com | 1 (33) |
| indiatoday.in | 115 (1,433) | tfc-taiwan.org.tw | 25 (1,077) | assamese.factcrescendo.com | 1 (40) |
| healthfeedback.org | 111 (328) | factcheckkorea.afp.com | 24 (194) | tamil.newschecker.in | 1 (26) |
| hindi.boomlive.in | 109 (1,372) | malumatfurus.org | 24 (731) | | |
| cekfakta.tempo.co | 95 (1,142) | rappler.com | 24 (350) | | |

**Table 12: The 70 languages queried, with the 41 languages in bold present in the final dataset.**

| | | | |
|---|---|---|---|
| Amharic | Georgian | Lithuanian | Sinhala |
| **Arabic** | **German** | **Macedonian** | **Slovak** |
| Armenian | **Greek** | **Malayalam** | Slovenian |
| Azerbaijani | **Gujarati** | **Malay** | **Spanish** |
| Basque | Haitian Creole | **Marathi** | Swedish |
| **Bengali** | **Hebrew** | Nepali | Tagalog |
| Bosnian | **Hindi** | **Norwegian** | **Tamil** |
| Bulgarian | **Hungarian** | **Oriya** | **Telugu** |
| **Burmese** | Icelandic | **Panjabi** | **Thai** |
| **Croatian** | **Indonesian** | Pashto | **Traditional Chinese** |
| Catalan | **Italian** | **Persian** | **Turkish** |
| **Czech** | **Japanese** | **Polish** | Ukranian |
| **Danish** | Kannada | **Portuguese** | **Urdu** |
| **Dutch** | **Kazakh** | Romanian | Uyghur |
| **English** | Khmer | **Russian** | Vietnamese |
| Estonian | **Korean** | **Serbian** | Welsh |
| **Filipino** | Lao | **Simplified Chinese** | |
| Finnish | Latvian | Sindhi | |
| **French** | | | |