MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset

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.

KEYWORDS

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

1 INTRODUCTION

While it may be possible to track the history of misinformation, or 'fake news', back to Octavian of the Roman Republic [43], or Browne in the 17th century [7], 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.

A universal consensus has yet to be reached on the definition of misinformation. One convincing definition of misinformation is that it is 'false or misleading' information, with a subcategory of misinformation being disinformation, which is misinformation with the intention to deceive [20]. In this work, we do not specifically distinguish disinformation from misinformation.

There exist over one hundred fact checking organisations that manually verify the veracity of claims made online, often within news articles or social media posts. This is a time consuming task involving a multitude of different documents and sources in order to classify a claim. To build intelligent tools to help with this process, datasets that accurately represent misinformation are required.

Online misinformation is multimodal, multilingual and multi-topical. The multimodal aspect of misinformation manifests online in the use of image and video in addition to the commonly studied textual communication. Moreover, we also posit that an additional modality is the social behaviour of users online, which exhibits itself in the form of a graph, or network, of interactions and behaviours. These interactions vary by platform, but on Twitter, they can be considered actions such as 'retweeting', 'quote tweeting' or 'replying' to a tweet, or 'following' a user. While research typically studies such modalities in isolation, or occasionally, some subset, the integration of all modalities may be necessary to accurately capture the underlying classification of misinformation.

The multilingual dimension of misinformation can be challenging due to the focus of existing research on misinformation within the English language, with most existing misinformation datasets only covering the English language. See Table 2 for an overview of existing misinformation datasets. Further, there exists a focus of natural language processing research towards the English language. Indeed, in Sections 3.1.2 and 3.3 we observe that the transformer based models perform better when text is translated to English when compared to a multilingual model.

Finally, misinformation has the potential to permeate across all aspects of society and life, and is not restricted to a single topic or domain. For example, a significant number of tweeted articles analysed on Twitter, in the months preceding the 2016 United States presidential election, contained fake or extremely biased news [6]. During the COVID-19 pandemic, the World Health Organization director-general stated that not only are they fighting the COVID-19 pandemic but also an 'infodemic' [38]. Naming but two examples, this alone provides further motivation that automated misinformation detection systems must not be trained on a single topic (e.g., COVID-19) or domain (e.g., politics), and thus justifies the collection of datasets that capture the pervasiveness of misinformation across many aspects of society.

Given the severity of online misinformation, there have been numerous public datasets made available for researchers to develop and evaluate automated misinformation detection models with. These publicly available datasets cover topics ranging from celebrities [32], rumours [46], politics [36] and health [23]. These datasets typically include data from a social network, usually Twitter, along with labels assigning them to a category, categorizing

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.

Given the short message length of posts on Twitter, we believe that additional context is required in order to properly capture how claims are being discussed on social networks. This can take the form of the media involved in the posts, articles that users are sharing on the social network, or indeed the social network of the user themselves (i.e. who they follow, who follows them), as well as interactions with these posts, such as replies or retweets. Therefore, we semantically link fact-checked claims not only to the social network posts, but also to this additional information. 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.

Further, most misinformation datasets consist of a single data dump which, given the dynamic nature of the problem, means that datasets can become outdated. Therefore, we open source our data collection and linking infrastructure which connects claims to social networks, `MuMiN-trawl`, in order to provide a platform for future research to continue to build and extend our work.

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 6.2:

(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 `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 [41].
- 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 RELATED WORK

There is a number of publicly available datasets on the topic of misinformation. Some datasets have a narrow focus on topics, for example on politics [42], COVID-19 [9, 23], or celebrities [32], but others, such as the `PHEME5` [46] and `PHEME9` [19] datasets, have explicitly sought to include data from different events, although typically with much smaller numbers of events, claims and tweets than `MuMiN`.

A popular approach is to make extensive use of fact-checking websites in order to provide ground truth labels for misinformation datasets. Popular fact checking data sources include PolitiFact[3], which has been used by numerous datasets such as `Liar` [42], which has around twelve thousand facts of a political nature and associated short statements. Others have used PolitiFact instead with news articles, such as those from `FakeNewsNet` [36, 37]. A more recent dataset from `FakeNewsNet`, [36], links articles to claims, along with tweets sharing the referenced articles. In both cases, the number labelled news articles are, again, fewer than `MuMiN`, with 240 news articles labelled in [37] and 1,056 news articles labelled in [36].

Other work has sought to extend the number of fact-checking organisations used to construct datasets such as the `CoAID` [9] and `MM-COVID` [23] datasets, which contain claims from 6 and 97 fact checking organisations, respectively. Of the 97 used by `MM-COVID`, 96 of them come from the Poynter fact-checking network of fact-checking organisations. In total they have over 4,000 claims. `MM-COVID` is the first to address the monolingual problem with existing datasets by including data in 6 different languages, albeit on a single topic. This dataset is perhaps the most related work to ours in that it addresses several of the problems we outline. However, our dataset, `MuMiN`, is significantly larger in terms of the number of claims, tweets, languages and indeed topics, as we do not limit our dataset to COVID-19 related misinformation. Further, we use a more sophisticated procedure to link claims to tweets; the `MM-COVID` dataset links an article to social media posts by searching for the URL, title and first sentence of the article on Twitter, while our dataset performs linking based on cosine similarity between transformer based embeddings of the claims, tweets and articles.

One important aspect of this line of research to consider is around evidence based fact checking approaches. This line of research seeks to utilise available evidence, such as online news sources, Wikipedia, as well as social networks, in order to classify a claim as true or

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

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%

false. Research by Popat et al. [33] proposes a system to replace manual verification of claims with such a system. Using only the text of the claim, they retrieve online search results using the text, and use linguistic features of the resulting articles, as well as the reliability of the sources in order to classify a claim. To deal with the variability in labels given by fact checkers, such as 'mostly true' or 'partially false', they map 'mostly' to true, and ignore those of a 'partial' veracity.

Another weakness of these approaches are that they tend to be limited in the fact checkers that they utilise for their sources. The MultiFC dataset [1] seeks to address this by including claims from 26 fact checking sources, in English, producing a dataset of 34,918 claims. While they do include extra context and metadata, they do not include additional modalities (such as images), nor do they include social network data. Recent work by Hanselowski et al. [16] released a dataset, that while containing documents retrieved from different domains, such as blogs, social networks and news websites, as with other work in this area, theirs has a strong focus on text, overlooking the relevant and rich information contained in other modalities such as images, videos and social graphs of interaction. The same sentiment applies also to the very recent X-FACT dataset [14], that while multilingual (25 languages), contains only text data.

While significant attention has been paid to the use of fact checking organisations as a source of ground truth for claim veracity and verification, there has also been work studying artificial claims. One such dataset is FEVER [39] which consists of 185,445 claims created by manipulating sentences from Wikipedia, and then annotated manually into one of three categories, supported, refuted or not enough information. Also using Wikipedia is the HoVer dataset [17], which addresses a weakness of the FEVER dataset, in that to verify claims, often a single Wikipedia article is not enough, and often requires multiple sources, or 'hops'. In HoVer, claims can require evidence from up to four Wikipedia articles. However, HoVer is still a monolingual dataset with an emphasis on text data, differing significantly from MuMiN, which considers multiple modalities across multiple languages as important characteristics to consider for this problem.

See Table 2 for an overview of these datasets, which demonstrates the key differences between them.

3 DATASET CREATION

The dataset creation consists of two parts, one concerning the claims and their fact-checked verdicts, and the second part concerning the collection of the surrounding social context. As described in Section 1, this lends itself to two application tasks, the first being, given a claim, predict its veracity given the social context. The second being, given a Twitter post to be fact-checked and its social context, predict the veracity of the claim made in the Twitter post. 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.

3.1 Claims

3.1.1 Data Collection. For the collection of fact-checked claims we utilise the Google Fact Check Tools API[4], which is a resource that collects fact-checked claims from fact-checking organisations around the world. This API was also used in [35] 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 started by querying the API for the queries coronavirus and covid to ensure that we got results from active fact-checking organisations. To ensure language diversity, we used the Google Translation API[5] to translate the two queries into 70 languages(see the appendix for a list of all the languages). We then queried the Fact Check API for up to 1,000 fact-checked claims for each of the resulting 132 queries.

From the collected fact-checked claims, we collected all the fact-checking organisations involved, resulting in a list of 115 fact-checking organisations(see the appendix for a list of all the organisations). From this list, we scraped all the fact-checked claims from each of them, from the fact-checking organisation's inception up until present day. This resulted in 128,070 fact-checked claims.

3.1.2 Data Processing. The claim data collected from the procedure in Section 3.1.1 also included various metadata, and we extracted the following: (1) source: The source of the claim, which can be both names of people as well as generic descriptions such as "multiple social media posts"; (2) reviewer: The URL of the fact-checking website that reviewed the claim; (3) language: The language the claim was uttered or written in; (4) verdict: The fact-checking organisation's verdict; (5) date: The date the claim was uttered. If this date was not available then the date of the review was used. If neither of those two were available then we extracted a potential date from the URL of the fact-checking article using the regular expression `[0-9]{4}-[0-9]{2}-[0-9]{2}` [6]. Note, from this, we release only the date, keywords from the claim, the predicted verdict

[4]https://developers.google.com/fact-check/tools/api/reference/rest
[5]https://cloud.google.com/translate/docs/reference/rest/
[6]This regular expression matches four, two and two numbers, separated by dashes. Thus, 2020-01-30 would be matched, but 20-01-30 would not.

Table 2: 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 [42]	12,836		✓					✓		
Weibo [18]		9,528	✓		✓		✓	✓		
PHEME5 [46]		5,802	✓		✓					✓
FNN-BuzzFeed [37]	182		✓			✓		✓	✓	
FNN-PolitiFact17 [37]	240		✓			✓		✓	✓	
PHEME9 [19]		6,425	✓		✓					✓
Celebrity [32]	200		✓			✓				
FakeNewsAMT [32]	240				✓	✓				
FEVER [39]	185,445				✓					
AFCSDC [3]	422		✓		✓	✓				
UKP Snopes [16]	6,422		✓		✓	✓				
MultiFC [1]	34,918		✓		✓	✓				
HoVer [17]	26,000				✓					
FNN-PolitiFact20 [36]	1,056	564,129	✓			✓	✓	✓		✓
FNN-GossipCop [36]	22,140	1,396,548	✓			✓	✓	✓		✓
CoAID [9]	4,251	160,667	(✓)			✓				✓
MM-COVID [23]	11,565	105,300	(✓)	✓		✓	✓	✓	✓	✓
UPFD-POL [11]	314	40,740†	✓			✓		✓		✓†
UPFD-GOS [11]	5,464	308,798†	✓			✓		✓		✓†
X-FACT [14]	31,189		✓	✓	✓					
MuMiN	12,914	21,565,018	✓	✓	✓	✓	✓	✓	✓	✓

(using the verdict classification model described below) and the reviewer involved.

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. Aside from the simple classifications such as labelling "false" and "misleading" as `misinformation`, and labelling "true" and "correct" as `factual`, we adhered to the following labelling guidelines. In the cases where the claim was "mostly true", we decided to label these as `factual`. When the claim was deemed "half true" or "half false" we labelled these as `misinformation`, with the justification that a statement containing a significant part of false information should be deemed as being misleading. When there was no clear verdict then the verdict was labelled as `other`. This happens when the answer is not known, or when the verdict is merely discussing the issue rather than assessing the veracity of the claim. The claims with the `other` label were not included in the final dataset.

To be able to properly deal with the multilingual verdicts, we attempted two approaches: (1) Translate them into English and use a pre-trained English language model; (2) Embed them using a pre-trained multilingual language model.

For the first approach, we used the Google Cloud Translation API[7] to translate the verdicts into English and train a model to classify these translated verdicts. The verdict classifier is based on the `roberta-base` model [24], with an attached classification module. This classification module consists of 10% dropout, followed by a linear layer with hidden size 768, a `tanh` activation, another 10% dropout layer, and finally a projection down to the three classes.

The model was trained on the dataset further augmented with casing. Specifically, we converted all the verdicts in the training set to lower case, upper case, title case and first letter capitalised. This resulted in a training dataset of 10,000 verdicts. We manually labelled 1,000 further verdicts to use as the test dataset. These verdicts were not deduplicated, to ensure that their distribution matches the true one. The model was trained for 10 epochs using the `AdamW` optimizer [25] with a learning rate of 2e-5, and a batch size of 8.[8] The model achieved a macro-F1 score of 0.99 among the `misinformation` and `factual` classes, and 0.92 if the `other` class is included.

For the second multilingual approach, we augmented the original (multilingual) verdicts, both by translating all of the 2,500 unique verdicts into 65 languages, using the Google Cloud Translation API[9] as well as applying the casing augmentation as described above, as finetuning a multilingual model directly on the original verdicts resulted in poor performance for the minority languages. The resulting dataset consisted of roughly 5 million verdicts, and we finetuned the `xlm-roberta-base` model [8] for 4 epochs on the dataset with the same hyperparameters as the model trained on the English-only dataset. On the same test dataset of 1,000 multilingual verdicts, this multilingual model achieved a macro-average F1-score of 0.98 among the `misinformation` and `factual` classes, and 0.85 if the `other` class is included.

As the English-only model was marginally better than the multilingual model, we opted to use that in building the dataset. However,

[7] See https://cloud.google.com/translate/docs/reference/rest/.

[8] This was done using the `transformers` [44] and `PyTorch` [30] frameworks

[9] See https://cloud.google.com/translate/docs/reference/rest/.

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[10].

See Table 3 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.

Table 3: 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

3.2 Twitter

From the claims and verdicts, we next collected relevant social media data. This data was collected from Twitter[11] using their Academic Search API[12], where we aimed to collect as many relevant Twitter threads that shared and discussed content related the claims obtained through the method described in Section 3.1.1.

3.2.1 Data Collection. From each claim, we first extracted the top 5 keyphrases[13], with a keyphrase being either one or two words from the claim, whose associated sentence embedding has a large cosine similarity to the sentence embedding of the entire claim.

We then queried the Twitter Academic Search API for the first 100 results for each of the five keyphrases, where we imposed that the tweets should not be replies[14], they had to share either a link or an image[15] and they had to have been posted at most three days prior to the claim date and at the latest on the claim date itself. The idea behind this is to obtain as high a recall as possible, i.e. obtaining as many of the potentially relevant tweets as possible, from which we can filter the irrelevant tweets. From the Twitter API we requested the tweets, users as well as media shared in the tweets. This resulted in approximately 30 million tweets.

3.2.2 Data Processing. As our goal with a automatic misinformation detection system is to be able to act on stories shared on social media before they go viral, we decided to filter the tweets to only keep the ones that have gained a reasonable amount of interest. We measure such interest in terms of 'retweets', and we chose a minimum of 5 retweets to be a conservative threshold, which reduced the number of tweets by 90%, leaving about 2.5 million tweets.

From these we then extracted all the URLs and hashtags from the tweets, as well as from the descriptions of each user. If the URL pointed to an article, we downloaded the title, body and top image[16]. We also extracted the hyperlinks of all images shared by the tweets. All of this information was then populated in a graph database[17] with approximately 17 million nodes and 50 million relations. Uniqueness constraints were imposed on all nodes.

3.3 Data Linking

From the database of tweets, the next task was to find all the Twitter threads that were relevant to each claim. As the tweets, claims and articles were multilingual, we again had to make the same choice as in Section 3.1.2: either embed the texts with a multilingual language model, or translate them to English and use an English language model. We did both, with the translation being vastly superior, as the multilingual model seemed to "collapse" on texts in certain languages like Thai, Tamil, Telugu, Bengali and Malayalam. Translating texts always comes with a risk of losing context, but as our goal was to find tweets that were discussing a claim at hand, we argue that a translated text will still be able to carry that information. The translation was done with the Google Translation API, as with the verdicts in Section 3.1.1.

Prior to embedding the approximately one million articles we first summarised the concatenated title and abstract, using the `BART-large-CNN` transformer [21]. This was done to enable embedding of additional tokens from the article content, as the Sentence Transformers have a limit of 512 tokens with the summarisation model being able to process 1,024 tokens. We then embedded these summarised articles along with the claims and tweets, all using the `paraphrase-mpnet-base-v2` Sentence Transformer [34].

Computing cosine similarities between every tweet-claim and article-claim pair would be computationally unfeasible. Instead, we grouped the claims in batches of size 100, fetched all the tweets and articles that were posted from three days prior to one of the claims in the batch up until the claim date, and computed cosine similarities between these[18].

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 Table 4. 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.

3.4 Data Enrichment

From the resulting Twitter posts linked to the claims as described in Section 3.3, we next queried Twitter for the surrounding context of these posts. We retrieved a sample of 100 users that retweeted

[10] See https://hf.co/saattrupdan/verdict-classifier-en and https://hf.co/saattrupdan/verdict-classifier.
[11] https://www.twitter.com
[12] https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all
[13] This was done using the KeyBERT [13] package together with the `paraphrase-multilingual-MiniLM-L12-v2` model from the Sentence Transformer package [34].
[14] Imposed with the query `-(is:reply OR is:retweet OR is:quote)`.
[15] Imposed with the query `(has:media OR has:links OR has:images)`.
[16] This was done using the `newspaper3k` Python library [29].
[17] We used the Neo4j framework, see https://neo4j.com/.
[18] This was done using the `PyTorch` framework [30].

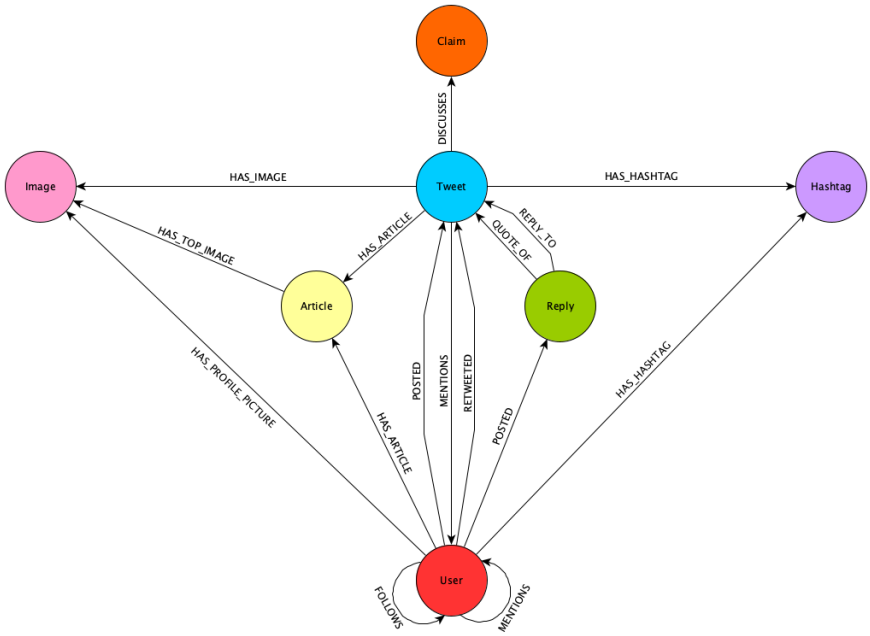

Figure 1: The graph schema of the MuMiN dataset.

Table 4: 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%

the tweets, 100 users who followed the authors of the tweets, 100 users who were followed by the authors of the tweets, 500 users who posted a reply to the tweets and all users who was mentioned in the tweets. For each of these users, we queried Twitter for their recent 100 tweets.

4 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 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 3.3, 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 [40]. A common way to fix this issue [9, 23] 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[19].

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 5). Further, to comply with copyright restrictions of the fact-checking websites, we do not release the claims themselves. Instead, we release keyphrases, obtained as described in Section 4.1. The datasets thus contain the tweet IDs, user IDs and claim keywords, as well as the `POSTED`, `MENTIONS`, `FOLLOWS`, `DISCUSSES` and `IS_REPLY_TO` relations, shown in Figure 1. From these, the remaining part of the dataset can be built by using our `mumin` package, see Section 5.

4.1 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` [28] projection of embeddings of the claims and `HDBSCAN` [26], a hierarchical density based clustering algorithm, we were able to discover 26 clusters based on the claim text. We optimized the hyperparameters of the projection as well as clustering algorithms[20], achieving a silhouette coefficient of 0.28. The clusters can be seen in Figure 2.

To provide context for each cluster, we concatenated the claims in each cluster and extracted keyphrases from each cluster[21]. 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.

5 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[22].

[19]This can be found at https://mumin-dataset.github.io/.

[20]This optimization 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` [31] and `hdbscan` [27].

[21]This was done using the KeyBERT library [13] on embeddings produced by a Sentence Transformer `paraphrase-multilingual-MiniLM-L12-v2` [34].

[22]See https://mumin-build.readthedocs.io for a full list of arguments.

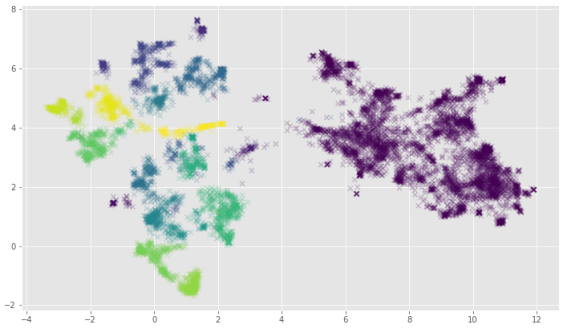

Figure 2: UMAP projection of the claim text embeddings. The large cluster on the right corresponds to COVID-19 related claims.

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 [41].

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[23].

6 MODEL PERFORMANCE

6.1 Dataset Splits

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 4.1. This is done as follows. We start by sorting all the claim clusters by size, in ascending fashion. We next add clusters into the validation set until at least 10% of the claims have been included. We then add clusters into the test set until at least 10% of the claims have been included, and the remaining clusters constitutes the training set. Statistics for each of the splits can be found in Table 5, which shows that we still roughly maintain the label balance throughout all the dataset splits.

6.2 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

[23]The tutorial and all the source code can be accessed through https://mumin-dataset.github.io/.

Table 5: 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

claim whose verdict is misinformation or factual. We name this task "tweet classification".

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 [12] with a linear classification head, and apply this model directly to the claims and the source tweets, respectively. We also benchmark a version of this model where the transformer layers are frozen, and name this model LaBSE-frozen. Secondly, we implement the vision transformer (ViT) model from [10], 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 [15], as follows. For each node, 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)[24]. We call this baseline model HeteroGraphSAGE.

See Table 6 and 7 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 [22, 45].

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.[25].

Table 6: 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-frozen	57.50%	54.10%	55.00%
LaBSE	62.55%	55.85%	57.90%
HeteroGraphSAGE	57.95%	57.70%	**59.80%**

Table 7: 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%**

7 DISCUSSION

7.1 Representative Data Splits

In the field of automatic misinformation detection, the splitting of the dataset into training/validation/test datasets is usually done uniformly at random [11, 23, 36, 39, 42, 46]. However, we argue that the main purpose of such a system is to be able to handle *new* events in which misinformation occurs, and therefore our dataset splits should reflect this. We conduct an experiment in which we analyse the performance differences of the baseline models if we had split the MuMiN dataset at random.

Concretely, we repeat two of our baselines on the random splits: the LaBSE classifier and the HeteroGraphSAGE model. Call the resulting models LaBSE-random and HeteroGraphSAGE-random.

On the tweet classification task, the LaBSE-random model achieved a macro-average F1-score of 71.10% and 73.6% on MuMiN-small and MuMiN-medium, respectively. The HeteroGraphSAGE-random model achieved 74.90%, 89.50% and 79.80% on MuMiN-small, MuMiN-medium and MuMiN-large, respectively. We see that the scores are drastically higher for these "random models" on this task compared to the results of the baseline models, as can be seen in Table 7.

For the claim classification task, the LaBSE-random model achieved a macro-average F1-score of 58.85%, 62.80% and 61.50% on MuMiN-small, MuMiN-medium and MuMiN-large, respectively. On this task, the HeteroGraphSAGE-random model achieved 48.50%, 61.40% and 62.55%, respectively. There is not as big of a difference between these "random models" and our baselines as with the tweet

[24]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.
[25]See https://mumin-dataset.github.io/ for both the leaderboard and the baseline repository.

classification task, as can be seen from Table 6, but the results are still marginally better than the baseline models.

This shows that having realistic splits of the dataset is important to guide our algorithm development in the right direction, especially in the field of automated misinformation detection, where we are interested in generalisability to new real-world events.

7.2 Limitations

Due to the automated linking procedure between facts and tweets, and facts and articles, erroneous labels exist. Nonetheless, this can be somewhat addressed by selecting higher similarity thresholds, as can be seen in Table 4. We did not make any judgements as to the impartiality or correctness of any of the verdicts provided by the fact-checking organisations. Therefore, this dataset may contain verdicts (labels) that are contentious or inaccurate. As a potential remedy, we provide the fact-checking organisation responsible for each claim and verdict. As the verdicts of each claim from fact-checking organisations are provided in unstructured freetext, we resorted to a machine learning model to classify each verdict into one of three categories, `factual`, `misinformation` or `other`. While we obtained a high performance on a test set, it is likely some verdicts may have been misclassified. As we do not distribute raw social network data, but instead provide code to retrieve it, this means that the dataset is truly dynamic such that if a user deletes a tweet or their account, their data will not be retrievable from the Twitter API. This makes reproducible research, if it involves the contents of the social network data, challenging.

7.3 Ethical Considerations

It is accepted that there are online harms associated with misinformation. Unfortunately, when the number of posts made on social networks daily is considered, the problem exists at a scale where manual curation is exceptionally difficult, thus motivating the use of automated methods to assist in the detection of misinformation online. These methods tend to utilise machine learning, and therefore typically require the collection of large amounts of data upon which to train the model. While the goal with such data collection is to combat an online harm, there is, understandably, ethical considerations related to the potential harms caused from the collection and use of large online datasets of social data, text data, and media data. A major factor for consideration is with respect to the collection of social network data, and the fact that this data is generated from users of the social network. The data collected in this dataset consists of only public Twitter data, accessed through the official Twitter Academic API. While the users of Twitter, in making their posts public, may expect their posts to be visible, in accordance with the Twitter developer terms, we do not include the raw collected data. Instead, we make available only tweet and user IDs, with associated code to 'hydrate' them (i.e. retrieve the full tweet and user data). Therefore, if a user deletes a tweet, or deletes their account, it will be no longer possible to retrieve the deleted data from what we have released. Thus, while we expect that data may disappear over time as a result, this trade-off is required. The ethics of this work has been approved, both by the University of Bristol Faculty of Engineering Research Ethics Committee (ref: 116665), as well as by the Ethics Board at the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN).

8 LICENSES

We release the three versions of the `MuMiN` under the Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0). The code, which includes the `mumin` package, the data collection and linking system `MuMiN-trawl`, as well as the repository containing the baselines, are all released under the MIT license.[26]

9 CONCLUSION

In this paper we presented `MuMiN`, which consists of a large scale graph misinformation dataset that contains 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 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 the `MuMiN` as well as providing easy export to Python graph machine learning libraries. Finally, we proposed and provided baseline results for two node classification tasks; a) predicting the veracity of a claim from its surrounding social context, and b) predicting the likelihood that a tweet to be fact-checked discusses a misleading claim. 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.

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*.

[26] See https://mumin-dataset.github.io/ for both `mumin`, `MuMiN-trawl` and the baselines.

[6] Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (2019), 7. https://doi.org/10.1038/s41467-018-07761-2

[7] 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.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

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

[10] 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).

[11] 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).

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

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

[14] 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

[15] 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.

[16] 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

[17] 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)*.

[18] 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.

[19] 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.

[20] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. https://doi.org/10.1126/science.aao2998 arXiv:https://science.sciencemag.org/content/359/6380/1094.full.pdf

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ACL* (2020).

[22] 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*.

[23] 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.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[25] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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

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

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

[29] Lucas Ou-Yang. 2018. Newspaper3k. *GitHub. Note: https://github.com/codelucas/newspaper, version 0.2.8.* (2018).

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[31] 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.

[32] 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.

[33] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 2173–2178. https://doi.org/10.1145/2983323.2983661

[34] 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

[35] 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).

[36] 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.

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

[38] The Lancet Infectious Diseases. 2020. The COVID-19 infodemic. *The Lancet Infectious Diseases* 20, 8 (2020), 875. https://doi.org/10.1016/S1473-3099(20)30565-X

[39] 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.

[40] 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

[41] 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

[42] 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.

[43] 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.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[45] 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.

[46] 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

In Table 8 we show the 70 languages that were queried during our data collection, and the resulting 41 languages, in bold, that made it to the final dataset. In Table 12 we show the 115 fact checking organisations whose fact checked claims have made it into the dataset. In Table 9, Table 10 and Table 11 we show language statistics across the large, medium and small versions of the dataset, respectively. In Figure 3 we show the cosine similarity distribution of the tweet-claim pairs.

Table 8: 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			

Table 9: 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 10: 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 11: 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%

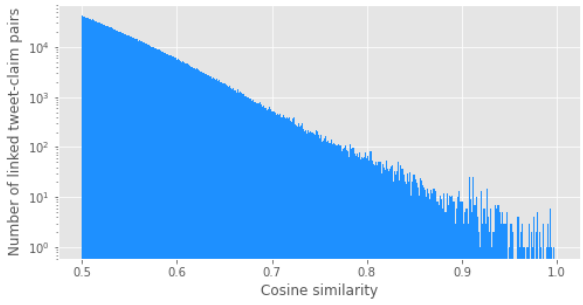

Figure 3: The distribution of cosine similarities among tweet-claim pairs.

Table 12: 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)		