Categorising Fine-to-Coarse Grained Misinformation: An Empirical Study of the COVID-19 Infodemic

Ye Jiang
University of Sheffield
Xingyi Song (x.song@sheffield.ac.uk)
University of Sheffield
Carolina Scarton
University of Sheffield
Iknoor Singh
University of Sheffield
Ahmet Aker
University of Sheffield
Kalina Bontcheva
University of Sheffield

Article

Keywords:

Posted Date: May 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1533519/v1

Additional Declarations: No competing interests reported.

Categorising Fine-to-Coarse Grained Misinformation: An Empirical Study of the COVID-19 Infodemic

Ye Jiang[1], **Xingyi Song**[1,*], **Carolina Scarton**[1], **Iknoor Singh**[1], **Ahmet Aker**[1,2], **and Kalina Bontcheva**[1]

[1]University of Sheffield, Department of Computer Science, Sheffield, S1 4DP, UK
[2]University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, 47057, Germany
[*]Corresponding author: x.song@sheffield.ac.uk

ABSTRACT

The spread of COVID-19 misinformation on social media has become a major challenge for citizens, with negative real-life consequences. Prior research has focused on detection and/or analysis of COVID-19 misinformation. However, finer-grained classification of misinformation claims has been largely overlooked. The novel contribution of this paper is in introducing a fine-grained annotated misinformation dataset which distinguishes between statements that assert, comment or question on false COVID-19 claims. This new dataset not only enables social behaviour analysis but also enables us to address both an evidence-based and non-evidence-based misinformation classification. Lastly, through a 'leave claim out' validation, we demonstrate that classifier performance on unseen COVID-19 misinformation claims is significantly different, as compared to performance on topics present in the training data.

1 Introduction

For a majority of citizens, social media became the primary source of information during the COVID-19 pandemic[1,2]. While social media allowed citizens to seek information in a more timely manner, it also resulted into an 'infodemic'[3] of misinformation which has caused significant harms. These have included death of more than 700 people from drinking denatured alcohol[4]; doctors being attacked due to misinformation such as "health workers were forcibly taking away Muslims and injecting them with the coronavirus"[5]; and 5G towers being burned down due to false claims that they caused COVID-19[6].

Even though international fact-checking outlets have increased 400% since 2014 in 60 countries[7], they alone are not able to eradicate COVID-19 false claims and online misinformation, in part due to the latter's sheer volume and velocity[1] but also due to their limited ability to reach the intended audience[8] and the ineffective promotion of reliable information by the platforms' algorithms. For instance, websites spreading misinformation had almost four times as many estimated views as equivalent content from reputable organisations on Facebook (https://secure.avaaz.org/campaign/en/facebook_threat_health/).

Therefore, while independent fact checkers (e.g., International Fact-Checking Network (IFCN) – https://www.poynter.org/ifcn/) play a vital role, they are increasingly reaching towards AI algorithms[9] to help scale up and optimise the fact-checking workflows. Such algorithms and models, however, have been developed primarily on datasets of political and other non-COVID-19 misinformation, which has impacted their accuracy in detecting and classifying COVID false claims.

Consequently, prior studies of COVID-19 misinformation focused mainly on misinformation detection[10–13], the social engagement with fake news on websites and social platforms[10], and the ways that misinformation is countered in tweets[8]. However, they have largely overlooked the wider online debates surrounding COVID-19 misinformation, such as the conversational threads around false COVID-19 claims and the questions and comments made as part of these. However, it is absolutely crucial for fact-checkers to have at their disposal algorithms that not only flag misinformation, but can also flag the comments and questions raised in online debates around false COVID-19 claims.

Consequently, this paper aims to address the following three research questions:

1. Which social media posts are propagating, questioning or commenting about a false claim?
2. Does the volume of tweets debunking a misinformation claim correlate with the volume of misinformation tweets?
3. What are the different kinds of misinformation spreading online?

In particular, the paper's novel contribution is in:

1. The creation of a large, unannotated dataset of COVID-19 tweets that are discussing IFCN fact-checked misinformation. In particular, these false claims are used as the queries to extract tweets with topics that are related to the particular false claim.

2. A manually annotated fine-grained COVID-19 misinformation dataset with 8 fine-grained categories that are suitable for training machine learning classification models for misinformation analysis.

3. A quantitative analysis of the fine-grained categories throughout a 10-month period and particularly investigating different kinds of misinformation tweets.

4. A benchmark experiment evaluating the performance of misinformation classifiers based on state-of-the-art Natural Language Processing (NLP) models on the 8 fine-grained categories. Specifically, the fine-grained classification enables not only the identification of misinformation, but also of related statements in the online discussion threads, including debunks, questions and comments.

5. The corresponding coarse-grained classification as (a) **Evidence based misinformation classification task** and (b) **Non-evidence based misinformation classification task**. In the first task, we aim to detect the misinformation that has already been debunked (the debunked misinformation that is provided by IFCN as the evidence). The misinformation prediction must be supported with verified misinformation. In the non-evidence based task, we aim to find social media posts that are likely to be misinformation; however these posts may require human verification.

Our earlier study[14] found that the topics of COVID-19 misinformation changed significantly throughout the different stages of the pandemic. Therefore, it is essential to evaluate the performance of misinformation detection classifiers on unseen topics as an indicator of their robustness and generalisability to new real-world data. To this end, we perform a 'leave false claim out' cross-validation (CV) to ensure that there is no topical overlaps between our training and testing data and compare performance against the standard random cross-validation approach.

1.1 Ethics Statement

The experiment processes undertaken as part of the SoBigData project have received ethical clearance from the University of Sheffield Ethics Board No. 025371. All experiments were performed in accordance with relevant guidelines and regulations.

2 Related Work

Research on misinformation prevalence, detection, and mitigation has become the focus of many research studies lately[15,16]. In addition, the dissemination of misinformation related to civil discourse[17], natural disasters[18] and health emergencies[2] has also been studied. Given the global popularity and easily accessible data on Twitter, past research has highlighted the importance of studying Twitter during epidemics. For example, 255 million Twitter users were active in February 2014 at the start of the Ebola outbreak[19] and this number topped 330 million in 2019[20]. Therefore, Twitter has become a rich source in studying the prevalence of online misinformation during the COVID-19 pandemic.

2.1 Claim Matching and Automated Fact Checking

There has been rigorous research in the development of automated fact-checking systems which includes multiple verification stages[9]. As proposed in CLEF-2020 and CLEF-2021 CheckThat! Lab task[21,22], claim matching is one of the pivotal stages during automated claim verification to find previously fact-checked claims. Moreover, recent study[23] shows existence of multiple false claims related to COVID-19 that spread in multiple modalities and languages on digital platforms. Fullfact[1] defines "claim matching" as the task of identifying the truth condition of unchecked claim using a set of fact-checked claims. Shaar et al.[24] proposed the task of detecting previously fact-checked claims as well as released a dataset of false claims along with their corresponding debunks from PolitiFact and Snopes. They formulated it as information retrieval task where the false statement from social media is used as a query to a corpus of fact-checked articles (around 16K). However, in this paper, we do exactly opposite where we use debunked claims as query to millions of tweets in order to find relevant tweet matches which includes misinformation, debunk, question etc (see Section 3.3 and Section 3.4 for more details). We further use this data to train misinformation classifiers based on state-of-the-art NLP models on the eight different fine-grained categories (see Section 4).

[1] https://fullfact.org/blog/2021/oct/towards-common-definition-claim-matching/

2.2 COVID-19 Datasets

With the outbreak of COVID-19, several datasets have been established to assist research communities to fight the pandemic. Singh et al.[25] investigate the early conversations about the pandemic on Twitter, and analyse five predefined myths as well as links to poor quality tweets between January and March 2020. Dong et al.[26] establish a real-time tracking of COVID-19 to help epidemiological forecasting. Chen et al.[27] collect COVID-19 scholarly articles for literature-based discoveries, and track the information spread on Twitter. To analyse how social behaviours are affected by the outbreak of COVID-19 and the spread of related information on social media on a large scale, Lamsal[28] collected 310 million English language tweets related to COVID-19 and analysed the sentiment, relations between countries and hashtags. Gruzd and May[29] release a multi-lingual Twitter dataset with around 270 million tweets. Gupta et al.[11] retrieve over 132 million tweets from around 20 million unique user IDs, and investigate their latent topics, sentiment and emotions by applying topic models and sentiment analysis.

In terms of datasets that particularly focused on misinformation related to COVID-19, Micallef et al.[8] investigate the tendency of the misinformation and counter-misinformation (aka. debunks) tweets based on two predefined topics (i.e. Fake Cures and 5G Conspiracy Theories). These datasets focus on predefined topics and themes, but topics of COVID-19 misinformation are fast-evolving. To tackle this, Cui and Lee[10] (CoAID) combine news articles published by reliable media outlets to identify the misinformation on Twitter. Sharma et al.[1] label tweets as misinformation if the tweet shared any article or content posted from any of the misinformation sources compiled using the fact-checking sources. However, it is hard to measure the reliability of such data since there is no gold-standard annotation. Saakyan et al.[30] (COVIDFACT) introduced a 'Counter-Claim' algorithm that automatically generated false COVID-19 related claims based on the subreddit r/COVID19 discussion, and obtains a moderate agreement of 0.47 for contradictory claims between models and humans. In this paper, we use professional IFCN verified claim for misinformation classification, but we are planning to introduce automatic generated false claims as an additional source to speed up the misinformation debunking. Hossain et al.[13] (COVIDLIES) divide misinformation detection into two sub-tasks: 1) relevant tweets retrieval based on COVID-19 misconceptions, and 2) stance detection to identify whether the tweets agree or disagree with the misconceptions. Several automatic methods are evaluated based on a manually annotated dataset, but the data time span is restricted to only a one-month period (i.e. from March to April 2020), meaning the assessment of long term tendencies of misinformation is not possible. Compared with the above dataset, our dataset investigates a longer time span over 10 months, which cover tweets from the first and second wave of outbreaks in the US and UK. We also use debunks which were provided by the professional fact-checkers (which provide evidence for the misinformation tweets).

2.3 COVID-19 Misinformation Detection

Singh et al.[25] define five Coronavirus Common Myths based on some keywords searching on the websites. Then the misinformation is identified based on phrases and words in the tweets and from broad descriptions of the myths (taken from the original searches that described each myth). Sharma et al.[1] compile information from three fact-checking sources (i.e., Media Bias/Fact Check, NewsGuard and Zimdars) that provide journalistic analysis of low-quality news sources known to frequently publish unreliable and false information. Similarly, Zhou et al.[12] apply Media Bias/Fact Check and NewsGuard to filter out news sites that are reliable/unreliable, and track misinformation based on the URLs and user information in the tweets.

Several studies apply machine learning methods to model the semantic feature in the misinformation. Micallef et al.[8] train three one-vs-all Logistic Regression classifiers to automatically identify misinformation, counter-misinformation and irrelevant tweets respectively. Cui and Lee[10] evaluate the hierarchical attention network (HAN)[31] and its variant dEFEND[32] on the CoAID datasets. Song et al.[14] propose a classification-aware neural topic model for a COVID-19 disinformation category classification and topic discovery. Meanwhile, Li et al.[33] evaluate the pre-trained language model BERT[34] with ensemble techniques on the AAAI 2021 Shared Task: COVID-19 Fake News Detection in English. Hossain et al.[13] combine BERTScore[35] with Sentence BERT to identify tweets stance for COVID-19 related misconceptions. However, those misinformation detection methods do not evaluate the effectiveness of using debunk information provided by the professional fact-checkers. In this paper, we investigate how the debunks would potentially effect the misinformation detection performance.

3 Dataset and Annotation

The overall pipeline of dataset annotation is shown in Figure 1. In general, we first collect tweets based on a set of keywords and create a COVID-19 related tweets index into the Elasticsearch (https://www.elastic.co/elasticsearch/). We also create a fact-check dataset, which includes fact-checked misinformation and its meta data from the IFCN websites, and select 90 misinformation claims as the queries to retrieve related tweets from the index. In order to improve the relevance of tweets, similar to previous research[13], we implement a Transformer-based model to re-rank the retrieved tweets based on their semantic similarities. The re-ranked tweets are then annotated based on fine-grained categories, and the agreement rates between annotators are evaluated. Finally, several classification tasks are conducted based on different types of data validation methods.

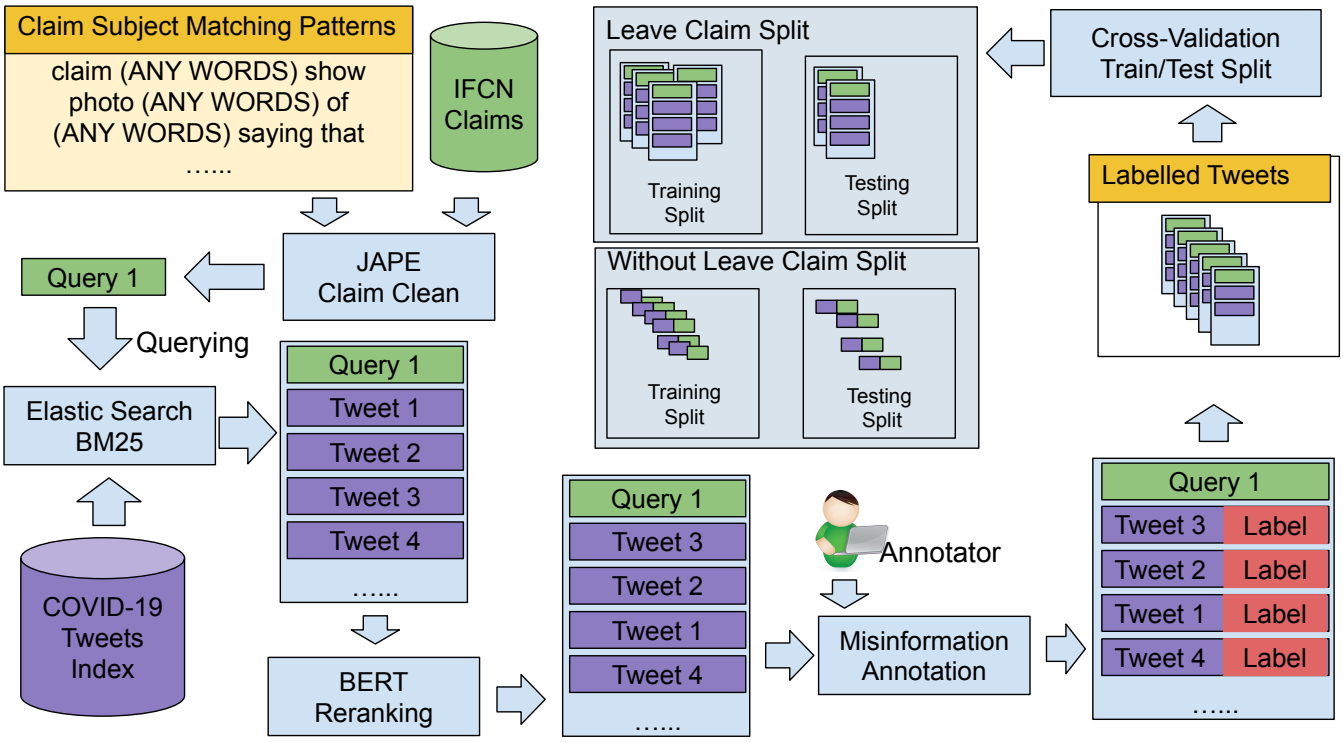


Figure 1. Overall pipeline

3.1 Tweet Collection

We first identify a collection of keywords (e.g, covid, covid-19, coronavirus, covid_19, etc.) related to COVID-19 and collect tweets that contain one of those keywords in the hashtag. We use the Twitter Stream API (https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data) to collect 182,027,646 English tweets spanning 10 months from March to December 2020. Then, we create ElasticSearch index for the tweets that are collected.

3.2 IFCN Dataset

In order to have a fact-checked list of COVID-19 related misinformation, we also build a IFCN dataset by utilising the work of fact-checkers. First, we extract 10,381 fact-checked misinformation claims (referred to as 'claims' in the remaining parts of the paper) from the IFCN Poynter website (https://www.poynter.org/ifcn-covid19-misinformation/). We select 90 English claims from April 2020, focusing on claims that appeared in the UK and US, since we wanted to maximise the number of tweets in English that could be retrieved. The IFCN claim extraction and process steps follow the same procedures as our previous research[14] A pattern matching language – JAPE[36] is applied to remove the subject from the claim in order to obtain a precise expression of the misinformation. e.g. "*Japanese doctor who won Nobel Prize said coronavirus is artificial and was manufactured in China*" the subject "Japanese doctor who won Nobel Prize said" is removed and the claim shortened to "*coronavirus is artificial and was manufactured in China*". The example subject patterns used in this work can be found in Figure 1 'Claim Subject Matching Patterns' (yellow) box.

3.3 Tweets Retrieval and Re-ranking

The selected 90 IFCN claims are used as the queries to retrieve tweets from the Elasticsearch index. Given the success of multistage neural re-ranking[37–39], we employ the same for retrieving relevant tweets. The initial retrieval utilises BM25 algorithm[40] to extract the 1,000 most relevant tweets from the Twitter index. To mitigate the cost of retrieval time, we then implement a tinyBERT[41] model, which has been pre-trained based on the MS MACRO dataset[42] for document ranking, to re-rank the retrieved tweets based on the semantic similarities between queries and tweets. After re-ranking, we select the 20 most relevant tweets for each misinformation, based on the cosine similarity scores. In addition, we restrict the retrieval for tweets posted in a date range of 10 weeks before and 2 weeks after the debunk date. This way, we aim to collect tweets related to a specific misinformation in a certain time, since similar misinformation can appear at different stages (e.g. misinformation about generic topics like 'a nurse in Italy died after taking the COVID-19 vaccine' may appear and re-appear at different times, in different countries, depending on the the vaccine roll out). Table 1 shows the results of our method for retrieving relevant

tweet matches. Here relevant tweet match can include a tweet which is a misinformation, related misinformation, debunk, related debunk, question or comment (Please refer Section 3.4 for the manual annotation process and further details of the classes). The results depict high retrieval performance with the mean average precision and precision of 0.88 and 0.93 for the top five retrieved tweets. Next, if we consider all the retrieved tweets, we achieve 0.89 mean average precision, demonstrating the effectiveness of our method for retrieving relevant tweet matches.

Metrics	**Mean Reciprocal Rank**	**Precision@K**				**Mean Average Precision@K**			
K	NA	1	5	10	All	1	5	10	All
Results	0.9401	0.9222	0.8844	0.8633	0.8400	0.9222	0.9312	0.9120	0.8902

Table 1. Tweet retrieval results

3.4 Annotation

The annotators carried out the work as part of their student research projects at the University of Duisburg-Essen and thus their informed consent was obtained verbally as part of enrolling to the project.

We obtained 1,800 tweets after the initial retrieval and re-ranking. Nine volunteer annotators were recruited and we gave them the instructions available in Appendix A for annotating tweets. The definition of fine-grained categories are listed as following:

1. **Misinformation**: Tweets contain falsehoods, inaccuracies, rumours, decontextualised truths, or misleading leaps of logic, and deliver exactly the SAME information/topic as the claim.
2. **Related Misinformation**: Tweets contain falsehoods, inaccuracies, rumours, decontextualised truths, or misleading leaps of logic, and deliver a SIMILAR information/topic with the claim but towards, for instance, a different person name, event name, medication name, illness name, etc.
3. **Debunk**: Tweets refute exactly the SAME information/topic as the claim, and are generated either by professional fact-checkers e.g.government website, IFCN, etc., or general citizen responses with/without use of any checkable evidence e.g. reputable links, hashtags, etc.
4. **Related Debunk**: Tweets refute a SIMILAR information/topic with the claim but towards, for instance, a different person name, event name, medication name, illness name, etc., and are generated either by professional fact-checkers e.g. government website, IFCN, etc., or general citizen responses with/without use of any checkable evidence e.g. reputable links, hashtags, etc.
5. **Question**: Tweets raise a question based on the exact SAME information/topic as the claim.
6. **Comments**: Tweets add some comments on the exact SAME information/topic as the claim.
7. **Relevant Others**: A tweet is not misinformation or a debunk of the claim but is nevertheless about the topic of the given claim.
8. **Irrelevant**: The information/topic of the Tweets that are IRRELEVANT to the claim.

Before the formal annotation, a pilot annotation was conducted so as to train the annotators. The formal annotation task was then conducted in a 3-weeks period. We created groups with three annotators each and we kept the same annotators in each group throughout the 3-weeks task, so each entry was annotated three times to evaluate the annotation agreements. Each annotator was assigned 200 tweets in each week.

During annotation, each entry provided to the annotators presented the query, the date when the misinformation was debunked, the fact-checkers' explanation, the organisation who fact-checked the misinformation, the misinformation veracity (e.g. false, misleading), and the source link to the fact-checkers' own web page. The volunteers assign each tweet with the most relevant of the eight fine-grained categories, and indicate their confidence (on a scale of 0 – least confident – to 5 – most confident) as well as their comments, if any. The tweet ID, the tweet text, the link to the tweet, and the date of when the tweet was posted were also provided.

We calculate the Krippendorff's alpha for each week to assess the data quality, and the final averaged score among the three weeks is 0.67, which demonstrates a substantial agreement between annotators. The final dataset is produced by merging the multiple-annotated tweets on the basis of: 1) majority agreement between the annotators where possible; or 2) confidence

score, if there was no majority agreement, the label with the highest confidence score was adopted. From the 1,800 tweets, 78 tweets did not have either majority agreement or a valid confidence score, so we removed those tweets in the final dataset. The statistics of the final annotated dataset are shown in Table 2 and examples of tweets in each class can be found in Appendix B.

Misinformation	Related Misinformation	Debunk	Related Debunk	Question
522	175	194	56	115
Comment	Irrelevant	Relevant Others	Total	
99	199	362	1722	

Table 2. Number of examples per category in the final dataset.

3.5 Data Analysis

One of the aims of this work is to understand the correlation of the spread of misinformation and debunk with other behaviours. As shown in Figure 2 (left), firstly, the volume of misinformation tweets is significantly higher than the other categories, especially at the beginning of April, which coincides with when the first wave of the pandemic started in both United State and United Kingdom. Secondly, there is a significantly higher volume of 'question and comment' tweets indicating that people tend to seek information and leave comments at the beginning of the first wave, but this tendency is decreasing throughout the pandemic. We also observe that there is a notable correlation between misinformation and debunk tweet counts (Pearson correlation $\rho = 0.55$, $p < 0.001$). This indicates that misinformation tweets and debunk tweets are spread at the same rate, similar to the findings made in[8] and[43]. The misinformation tweets also have a positive correlation with comment tweets (Pearson correlation $\rho = 0.58$, $p < 0.001$) and question tweets (Pearson correlation $\rho = 0.45$, $p < 0.001$), this is similar to the debunk tweets with comment tweets (Pearson correlation $\rho = 0.54$, $p < 0.001$) and question tweets (Pearson correlation $\rho = 0.41$, $p < 0.001$).

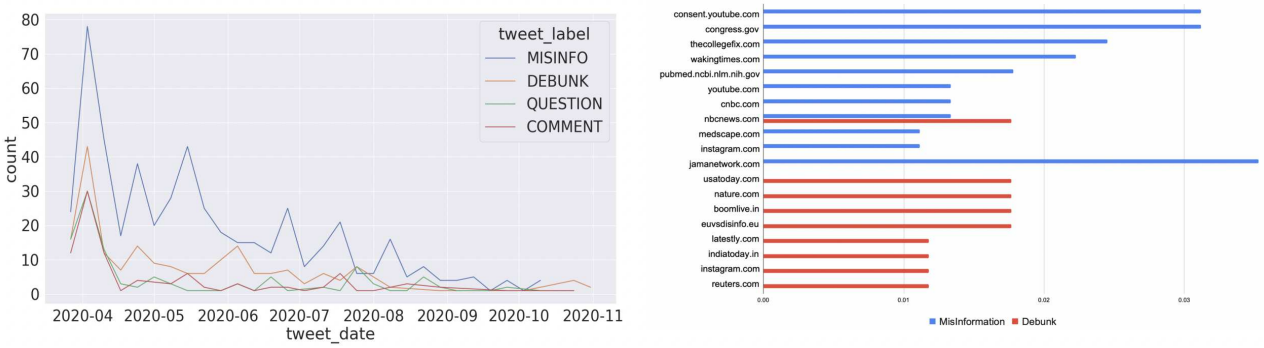


Figure 2. Left: Misinformation, debunk, question and comment tweets volume over time (in weeks). Right: Top 10 frequent URLs found in misinformation and debunk tweets.

Overall, we find that the debunk tweets have a similar spreading rate as misinformation tweets. In addition, people tend to leave comments or ask questions when there is a significantly high number of misinformation and debunks tweets.

3.5.1 URL Sources in Misinformation and Debunk Tweets

The top 10 frequent URL domain names found in misinformation and debunk tweets shown in Figure 2 (right). The numbers in horizontal axe are averaged by number of misinformation/debunk tweets. We note that there is almost no URL overlap between misinformation and debunk tweets (only overlap URL is cnbc.com), and misinformation tweets are very likely to link to video website (e.g. youtube.com). We also note that misinformation tweets have high frequency contain URLs than that in the debunk tweets, and may also contain high-credibility sources (e.g.PubMed). For instance, a misinformation tweet claims that *'Now officially : 5G Technology and induction of coronavirus in skin cells published online ahead of print, 2020 Jul 16. J Biol Regul Homeost Agents, 2020'* and provides a link to 'pubmed.ncbi.nlm.nih.gov'. However, that paper was retracted after a thorough investigation as it showed evidence of substantial manipulation of the peer review. In addition, several tweets quote information from *'clinicaltrials.gov'* and claim that *'Hydroxychloroquine and Zinc With Either Azithromycin or Doxycycline for Treatment of COVID-19 in Outpatient Setting'*. However, large-scale clinical trials demonstrate no beneficial effect of hydroxychloroquine in terms of viral shedding, disease severity, or mortality among COVID-19 patients.

3.5.2 Hashtags in Misinformation and Debunk Tweets

Similarly to URLs, we note that a 'hashtag' is a strong indicator to misinformation as well as debunk tweets. We found that some misinformation hashtags have negative emotion towards a person or an organisation (e.g., EvilGates, FireFauci, etc.) and some are generally denying the pandemic (e.g., FakePandemic, coronascam, etc.). On the other hand, hashtags in debunk tweets are less emotional (e.g., FactMatter, SeekReliableSource, etc.), and some directly indicate the professional fact-checkers or high-credibility source (e.g., AltNewsFactCheck, pubmed, PIBFactCheck, etc.). Wordclouds of misinformation and debunk tweets can be found in Appendix C.

4 Misinformation Classification Experiments

In this section, we conduct a benchmark experiment for our annotated Twitter misinformation classification dataset. This experiment includes three tasks that represent three different misinformation classification scenarios. The task detail and the experiment settings are discussed in Section 4.1. Then, we introduce the baseline models and model configurations in Section 4.2. Finally, the experimental results are discussed in Section 4.3.

4.1 Misinformation Classification Tasks

The classification experiment is divided into three tasks. Besides the fine-grained classification task, which takes account of all labels based on the evidences, we also introduce two coarse-grained classification tasks according to the different hierarchy methods of the fine-grained classes. The descriptions of each task are listed in the following paragraphs, and the corresponding labels for coarse-grained non-evidence based and evidence-based classification tasks are illustrated in Table 3.

Coarse-grained Evidence Based Classification		
Misinformation	**Debunk**	**Other**
Misinformation	Debunk	Comment Relevant Other Irrelevant Related Misinformation Question Related Debunk
Coarse-grained Non-Evidence Based Classification		
Misinformation	**Debunk**	**Other**
Misinformation Related Misinformation	Debunk Related Debunk	Question Comment Relevant Other Irrelevant

Table 3. Coarse-grained classification label hierarchy. The bold texts are the coarse-grained labels, and its corresponding fine-grained labels are listed in the column beneath

1. **Fine-grained misinformation classification**: Classify the tweet text into one of the eight fine-grained labels introduced in this paper. This task aims to identify the tweets that might be misinformation, debunk or other associated behaviours (e.g.tweets that leave comments about debunks or tweets that question about misinformation, etc). Since the information/topics of 'Misinformation' and 'Debunk' tweets are the same as the IFCN claim, and IFCN claims are served a evidences in our classification task, the fine-grained misinformation classification task is therefore evidence based.

2. **Coarse-grained evidence based misinformation classification**: Similar to fine-grained classification, this task aims to classify tweets that have already been debunked, but concentrates more on the misinformation and debunk tweets. In this case, tweets labelled with 'Misinformation' will be treated as '*Misinformation*' tweets and tweets labelled with 'Debunk' will be treated as '*Debunk*' misinformation. All other labels, including 'Related Misinformation/Debunk' are categorised as '*Other*'.

3. **Coarse-grained Non-evidence based misinformation classification**: This task aims to classify tweets likely to be misinformation, where there are no debunks available. Therefore, different to the coarse-grained evidence based task, the 'Related Misinformation/Debunk' labels are categorised as '*Misinformation/debunks*', together with 'Misinformation/Debunk' tweets.

For each classification task, we report the results based on 5-fold cross-validation. The evaluation metrics used in this experiments are 1) accuracy, 2) F1 measure for each class, and 3) macro average F1 (i.e. the average of class level F1 Measure) across all classes. Two different folding methods are used in this experiment:

- Folding **without Leave Claim Out**: This is the standard 5-fold cross-validation. The training data is randomly split into five sub-groups. For each sub-group, one sub-group is retained as the validation set, and the remaining sub-groups are used for training.
- Folding **with Leave Claim Out**: Similar to the standard 5-fold cross-validation, but the random sub-group splitting is based on claim rather than on all training data. Therefore no claim in the test set will appear in the training stage. This is a more realistic testing method to test model performance on 'unseen' misinformation since most of the online misinformation have not been debunked by the professional fact-checkers in the real world.

4.2 Model and Configuration

Four state-of-the-art baseline models are used in this experiment to benchmark the classification task performance. BERT_CLS and CANTM are the evidence independent models used to test the classification performance without providing claim (please note, claims are applied in this work as evidence) information. BERT_Pair and SBERT are evidence dependent models and have been widely applied in Natural Language Inference tasks. In this experiment, we apply these two models to test the performance with the aid of evidence information.

- *BERT_CLS*: The BERT[34] version used in this experiment is a 24 transformer layers (BERT-large) COVID-Twitter pre-trained[44] BERT. Only the parameters in the last transformer encoding layer is unlocked for fine-tuning, the rest of the BERT weights were frozen for this experiment. BERT_CLS treat all tasks as a Tweet text classification task. The model input is [CLS] + Tweet_Text + [SEP], and the final hidden state of [CLS] token will be the representation of Tweet_Text. The probability of labels is predicted using a Softmax classifier based on the Tweet_Text [CLS] representation.
- *CANTM*: Classification-Aware Neural Topic Model is a stacked asymmetric variational autoencoder that outputs classification and topic predictions. In this experiment, we only consider the classification output of CANTM model. CANTM apply the BERT model as input text encoder, and the BERT model setting is the same as BERT_CLS. The vocabulary size for CANTM is 3,000 with 50 latent topics.
- *Sentence-BERT* (SBERT): We apply SBERT[45] classification objective function for our classification experiment. SBERT classification objective function aiming to optimise the cross-entropy loss of a softmax classifier ($o = softmax(W(q,t,|q-t|))$). The input feature of the classifier is the weighted concatenation of evidence embedding (q), tweet text embedding (t) and the element-wise difference $|q-t|$. In this experiment, all embeddings are obtained from [CLS] token of COVID-Twitter pre-trained[44] BERT, and apply the same setting as *BERT_CLS*. The evidence of the tweet text is the claim that is described in Section 3.3.
- *BERT_Pair*: Similar to BERT_CLS, but BERT_Pair also takes evidence into consideration. The input of the model is [CLS] + Evidence + [SEP] + Tweet_Text + [SEP]. BERT_Pair has been originally applied for the next sentence prediction task and has been fine-tuned for pair-wise text classification such as Natural Language Inference. The probability of labels is predicted using a Softmax classifier based on the pairwise [CLS] representation. We experiment BERT_Pair model with two different settings: 1) The results labelled with BERT_Pair_MNLI are trained with the Multi-Genre Natural Language Inference (MNLI) corpus[46]. The MNLI labels contradiction, entailment and neutral corresponding to the debunk, misinformation, and other in our misinformation classification task. 2) The results labelled with BERT_Pair are trained with our labelled misinformation data (5-fold cross-validation)

4.3 Coarse-Grained Classification Results

Table 4 shows the results of coarse-grained misinformation classification tasks. In the without 'leave claim out' cross validation all models achieved more than 0.75 accuracy in both evidence- and non-evidence-based classification tasks. The best performed models are SBERT and BERT_Pair. Both models are evidence dependent and able to reach around 0.8 classification accuracy in both coarse-grained tasks.

Compared between two coarse-grained tasks, all baseline models have lower average F1 scores in the evidence-based classification task than non-evidence-based classification. This may be because: 1) *Evidence-based classification is a more challenging task*. In the non-evidence-based classification, the misinformation or debunks can be determined according to previously learned topics/information that was included in the training data. However, evidence-based classification is a pairwise classification task, misinformation/debunks can only be determined according to the given evidence. Hence, a tweet text cannot be classified as misinformation/debunk if it does not match the given evidence even the tweet text is misinformation/debunk

Without Leave Claim Out Cross Validation										
	Non-Evidence-Based Classification Task					Evidence-Based Classification Task				
	Acc.	Avg. F1	Debunk F1	MisInfo F1	Other F1	Acc	Avg. F1	Debunk F1	MisInfo F1	Other F1
BERT_CLS	0.789	0.771	0.709	0.803	0.799	0.759	0.715	0.608	0.729	0.808
CANTM	0.792	0.762	0.664	0.816	0.806	0.779	0.722	0.597	0.739	0.830
SBERT	0.808	0.789	0.724	0.815	0.828	0.804	0.753	0.643	0.765	0.851
BERT_Pair	0.797	0.787	0.749	0.807	0.804	0.808	0.757	0.665	0.760	0.846
With Leave Claim Out Cross Validation										
BERT_CLS	0.648	0.609	0.487	0.672	0.668	0.632	0.533	0.405	0.490	0.705
CANTM	0.640	0.584	0.448	0.647	0.657	0.622	0.477	0.252	0.453	0.724
SBERT	0.662	0.613	0.476	0.681	0.681	0.632	0.550	0.409	0.526	0.715
BERT_Pair	0.634	0.595	0.470	0.656	0.657	0.643	0.567	0.468	0.508	0.724
BERT_Pair_MNLI	0.455	0.396	0.384	0.227	0.578	0.514	0.395	0.312	0.219	0.655

Table 4. COVID-19 coarse-grained misinformation classification results.

(with other evidence). 2) *Data is more imbalanced in evidence-based classification task*. According to the label hierarchy (Table 3), related misinformation and debunks are categorised as 'Other' class in the evidence-based classification. This reduces the number of training samples in the misinformation/debunks classes, and increases the samples in the other class. According to the results, although the average F1 scores are lower, the 'Other' class F1 scores are better than the non-evidence-based classification task.

In the 'leave claim out' cross-validation, all models decreased at least 15% in average F1 measure compared to 'without leave claim out' cross-validation. This is expected, since in the 'leave claim out' cross-validation, the topics between training and testing set are different, and models cannot make a prediction based on its learned misinformation topics. According to the results, models are over-fitted to the misinformation topics from the training set. This also indicates that keeping the training data up-to-date is important to maintain the model's real-world misinformation classification performance.

According to the class-level F1 score, the performance of misinformation classification is better than debunk classification. This may happen because of the class imbalance problem. The number of debunk and related debunk samples is much smaller (about $1/3$) than misinformation and related misinformation samples. This problem is also reflected in the number of debunking posts being much smaller than the misinformation posts on social media. A faster misinformation debunk using an automated NLP algorithm will help prevent misinformation.

In the last row of Table 4 the classification performance of the Multi-Genre Natural Language Inference trained BERT_Pair$_{MNLI}$ model is shown (the average F1 score of MNLI mismatched development set is 0.73). The BERT_Pair$_{MNLI}$ have almost identical F1 score (0.39) in both tasks. Hence, the traditional natural language inference trained model may not be suitable for misinformation classification.

4.4 Fine-Grained Classification Results

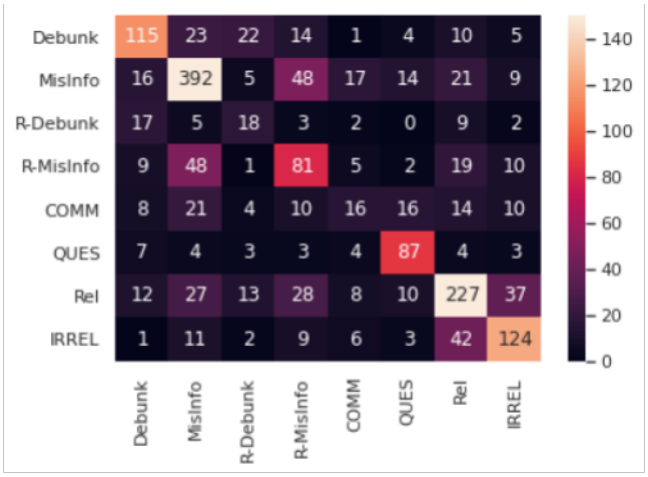
Table 5 shows the results of the fine-grained misinformation classification task. The fine-grained misinformation classification task is evidence-based. This task further split the *other* class from the coarse-grained evidence-based classification task into six more granular classes (Related Debunk, Related Misinformation, Comment, Question, Relevant Other and Irrelevant) according to the given evidence. In the 'without leave claim out' cross-validation, all models drop around 0.2 average F1 scores compared to the coarse-grained evidence-based classification task. The main performance decrease occurred in the fine-grained 'Other' classes. The debunk and misinformation class-level F1 measure remains similar in performance (but slightly worse) as the coarse-grained evidence-based classification task. This is because the number of misinformation and debunk training samples are the same as coarse-grained evidence-based classification. The main challenge of the fine-grained classification is to predict samples from 'Other' classes further into six fine-grained classes.

Figure 3 (a) shows the confusion matrix of BERT_Pair results in the fine-grained classification 'without leave claim out' validation. According to the figure, most 'Related Debunk/Misinformation' samples are misclassified as 'Debunk/Misinformation'. This may happen because all training samples are semantically similar to the IFCN claim (the training samples are the top 20 tweets with the highest BERT embedding cosine similarity to claim), and the model is unable to catch the difference between them. An example of this error type is presented in Figure 3 (b), Claim 1.

The misinformation claim states that steam from "boiling oranges" kills COVID-19. However, the tweet text being classified is debunking steam from "boiling water" kills COVID-19. The debunk is not directly addressing the query misinformation,

	Without Leave Claim				Leave Claim			
	BERT_CLS	CANTM	SBERT	BERT_Pair	BERT_CLS	CANTM	SBERT	BERT_Pair
Accuracy	0.584	0.621	0.639	0.615	0.310	0.349	0.353	0.370
F1	0.515	0.524	0.555	0.524	0.271	0.277	0.259	0.276
Debunk F1	0.622	0.638	0.630	0.602	0.333	0.312	0.361	0.382
MisInfo F1	0.671	0.736	0.757	0.742	0.373	0.476	0.535	0.495
R-Debunk F1	0.293	0.264	0.409	0.258	0.025	0.0	0.071	0.038
R-MisInfo F1	0.416	0.439	0.478	0.434	0.135	0.085	0.069	0.131
COMM F1	0.239	0.224	0.159	0.209	0.110	0.221	0.143	0.149
QUES F1	0.715	0.695	0.719	0.697	0.613	0.623	0.451	0.578
REL F1	0.595	0.624	0.646	0.635	0.335	0.343	0.309	0.320
IRREL F1	0.573	0.572	0.643	0.613	0.248	0.158	0.131	0.116

Table 5. COVID-19 misinformation fine-grained query based classification. The corresponding class label are R-Debunk:Related Debunk, R-MisInfo:Related Misinformation, COMM:comment, QUES:question, REL:Relevant Other, IRREL:irrelevant



(a)

Claim 1	Steam from boiling oranges kills COVID-19.
Tweet Text	#Fact: No scientific evidence to prove that inhaling hot water steam kills #Coronavirus
Prediction: DEBUNK	Label: RELATED_DEBUNK

Claim 2	Research proves that commercial mouthwash could protect against COVID-19.
Tweet Text	Mouthwash could prevent COVID-19 transmission, scientists say https://… via @... @... This is a reckless headline. It should read, "Scientists theorize mouthwash may prevent COVID-19, more research needed." #scicomm #covid19 cc @...
Prediction: MISINFORMATION	Label: COMMENT

(b)

Figure 3. (a) BERT_Pair confusion matrix in the fine-grained classification 'without leave claim out' validation. Numbers in each row are the number of samples labelled in the corresponding class, and numbers in each column are the number of samples which have been predicted in the corresponding class. (b) Example of misclassified cases.

therefore, the label should be "RELATED DEBUNK".

Another major classification error occurs in the 'Comment' class. The class level F1 scores for the 'Comment' class are less than 0.25 with all baseline models. According to the confusion matrix, the 'Comment' labelled samples are very likely to be classified as misinformation. The comment class contains tweets that make a comment about the misinformation. Therefore,

the misinformation is included in the comment tweet, which might be the main cause of this error. In Figure 3 (b), Claim 2 is an example of comment text. The tweet text quote a misinformation claim 'Mouthwash could prevent COVID-19 transmission' and make comment that 'more research needed' for this claim.

In 'leave claim out' cross-validation, all model average F1 score less than 0.3. Therefore, none of the baseline models are reliable for unseen fine-grained misinformation classification. This may be because all models are over-fitted with training data due to the limited number of samples in most classes. We also note that, only the 'Misinformation' class-level F1 score remains similar to the coarse-grained query-based task, and the 'Misinformation' class have the most number of samples in the dataset.

5 Conclusion

This paper introduced a fine-grained COVID-19 misinformation dataset, which contains 1,722 tweets with eight categories that are manually annotated. In our dataset, each tweet is triple annotated and the averaged Krippendorff's alpha is 0.67 which indicates a substantial agreement. To answer the research question above, we first found that misinformation tweets have similar spread rate to debunk tweets. Secondly, our dataset also enables the investigation of the occurrences of other social behaviours (e.g. questions or comments related to a misinformation) in tweets. We found both question and comment tweets have positive correlation with misinformation and debunk tweets. Thirdly, we also found that misinformation tweets can contain a URL from high-credibility sources. In addition, the hashtags in misinformation tweets are found to be more emotional, and debunk hashtags are more related to the professional fact-checkers. Our experiments in Section 4 conduct three misinformation classification benchmark experiments: 1) Non-evidence based classification 2) Evidence based classification and 3) Fine-grained classification. The results demonstrate that the all baseline models well performed in standard 'without leave claim out' validation across all classification tasks. However, the classification performance dropped significantly with 'leave claim out' setting. Therefore, we need to regularly update training instances to ensure the classification performance over time. In the future, we need to develop a classification method to adapt to the fast topic changing nature of misinformation.

6 Acknowledgements

This work is partially funded by the EU H2020 SoBigData++ (grant agreement: 871042) projects.

7 Data availability

The datasets generated and analysed during the current study are available in the Kaggle repository, https://www.kaggle.com/datasets/51f4a2c2d9e36ebc1e3e3411cb2fc5aeaf58db70fefb55a6191193ddd8eb4ae7. Please note, only Twitter id associated with the tweets data used in our study is publicly available. Users may require using Twitter API to reconstruct the full dataset. The permission to access Twitter API requires contacting Twitter directly.

References

1. Sharma, K., Seo, S., Meng, C., Rambhatla, S. & Liu, Y. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv e-prints* arXiv–2003 (2020).
2. Zhou, C., Xiu, H., Wang, Y. & Yu, X. Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on covid-19. *Inf. Process. & Manag.* **58**, 102554 (2021).
3. WHO. Novel coronavirus (2019-ncov). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf (2020).
4. Mehrpour, O. & Sadeghi, M. Toll of acute methanol poisoning for preventing covid-19. *Arch. toxicology* **94**, 2259–2260 (2020).
5. Khan, A. Indore stone pelting: The inside story of whatsapp messages and fear-mongering that led to shocking attack on doctors. https://www.freepressjournal.in/india/indore-stone-pelting-the-inside-story-of-whatsapp-messages-and-fearmongering-that-led-to-shocking-attack-on-doctors (2020).
6. BBC. Mast fire probe amid 5g coronavirus claims. https://www.bbc.co.uk/news/uk-england-52164358 (2020).
7. Stencel, M. Number of fact-checking outlets surges to 188 in more than 60 countries. https://www.poynter.org/fact-checking/2019/number-of-fact-checking-outlets-surges-to-188-in-more-than-60-countries/ (2020).
8. Micallef, N., He, B., Kumar, S., Ahamad, M. & Memon, N. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *arXiv preprint arXiv:2011.05773* (2020).

9. Zeng, X., Abumansour, A. S. & Zubiaga, A. Automated fact-checking: A survey. *Lang. Linguist. Compass* **15**, e12438 (2021).

10. Cui, L. & Lee, D. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).

11. Gupta, R. K., Vishwanath, A. & Yang, Y. Global reactions to covid-19 on twitter: A labelled dataset with latent topic, sentiment and emotion attributes (2021). 2007.06954.

12. Zhou, X., Mulay, A., Ferrara, E. & Zafarani, R. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3205–3212 (2020).

13. Hossain, T. *et al.* Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020* (2020).

14. Song, X. *et al.* Classification aware neural topic model for covid-19 disinformation categorisation. *PloS one* **16**, e0247086 (2021).

15. Sharma, K. *et al.* Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intell. Syst. Technol. (TIST)* **10**, 1–42 (2019).

16. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2018).

17. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. economic perspectives* **31**, 211–36 (2017).

18. Gupta, A., Lamba, H., Kumaraguru, P. & Joshi, A. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 729–736 (2013).

19. Twitter. Twitter reports first quarter 2014 results. https://s22.q4cdn.com/826641620/files/doc_financials/2014/q1/2014_Q1_Earnings_Release.pdf (2014).

20. Twitter. Twitter reports first quarter 2019 results. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Earnings-Release.pdf (2019).

21. Barrón-Cedeno, A. *et al.* Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 215–236 (Springer, 2020).

22. Nakov, P. *et al.* The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *ECIR (2)* (2021).

23. Singh, I., Bontcheva, K. & Scarton, C. The false covid-19 narratives that keep being debunked: A spatiotemporal analysis. *arXiv preprint arXiv:2107.12303* (2021).

24. Shaar, S., Babulkov, N., Da San Martino, G. & Nakov, P. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3607–3618, DOI: 10.18653/v1/2020.acl-main.332 (Association for Computational Linguistics, Online, 2020).

25. Singh, L. *et al.* A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907* (2020).

26. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases* **20**, 533–534 (2020).

27. Chen, E., Lerman, K. & Ferrara, E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Heal. Surveillance* **6**, e19273 (2020).

28. Lamsal, R. Coronavirus (covid-19) tweets dataset, DOI: 10.21227/781w-ef42 (2020).

29. Gruzd, A. & Mai, P. COVID-19 Twitter Dataset, DOI: 10.5683/SP2/PXF2CU (2020).

30. Saakyan, A., Chakrabarty, T. & Muresan, S. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794* (2021).

31. Yang, Z. *et al.* Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489 (2016).

32. Shu, K., Cui, L., Wang, S., Lee, D. & Liu, H. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405 (2019).

33. Li, X., Xia, Y., Long, X., Li, Z. & Li, S. Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english. *arXiv preprint arXiv:2101.02359* (2021).

34. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

35. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

36. Cunningham, H. *et al.* Jape: a java annotation patterns engine. In *Proceedings of the Workshop on Ontologies and Language Resources (OntoLex'2000)* (Department of Computer Science, University of Sheffield, 2000).

37. Nogueira, R. & Cho, K. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085* (2019).

38. Karpukhin, V. *et al.* Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781 (2020).

39. Singh, I., Scarton, C. & Bontcheva, K. Multistage bicross encoder for multilingual access to covid-19 health information. *Plos one* **16**, e0256874 (2021).

40. Robertson, S. E. *et al.* Okapi at trec-3. *Nist Special Publ. Sp* **109**, 109 (1995).

41. Jiao, X. *et al.* Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* (2019).

42. Nguyen, T. *et al.* Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS* (2016).

43. Mendoza, M., Poblete, B. & Castillo, C. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, 71–79 (2010).

44. Müller, M., Salathé, M. & Kummervold, P. E. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* (2020).

45. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3973–3983 (2019).

46. Williams, A., Nangia, N. & Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122 (Association for Computational Linguistics, 2018).

8 AUTHOR CONTRIBUTIONS STATEMENT

YJ, XS, and IS drafted the original manuscript. YJ, CS, AA and KB were involved in the data collection and annotation. YJ, IS conducted the Re-ranking experiment mentioned in Section 3. XS conducted the Misinformation Classification Experiments described in Section 4. KB and CS provided Project administration and supervision for this project. All authors reviewed the manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixA.pdf
- AppendixB.pdf
- AppendixC.pdf