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

Ye Jiang, Xingyi Song, Carolina Scarton, Iknoor Singh,
Ahmet Aker, Kalina Bontcheva
University of Sheffield, Sheffield, United Kingdom
{ye.jiang, x.song, c.scarton, i.singh,
ahmet.aker, k.bontcheva}@sheffield.ac.uk

Abstract

The spread of COVID-19 misinformation on social media became a major challenge for citizens, with negative real-life consequences. Prior research focused on detection and/or analysis of COVID-19 misinformation. However, fine-grained classification of misinformation claims has been largely overlooked. The novel contribution of this paper is in introducing a new dataset[1] which makes fine-grained distinctions 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 evidence-based and non-evidence-based misinformation classification tasks. Lastly, through *leave claim out* cross-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 the majority of citizens, social media became the primary source of information during the COVID-19 pandemic (Sharma et al., 2020; Zhou et al., 2021). While social media allowed citizens to seek information in a more timely manner, it also resulted in an 'infodemic' (WHO, 2020) of misinformation which has caused significant harms.

Therefore, while independent fact-checkers (e.g., International Fact-Checking Network IFCN[2]) played a vital role, they increasingly need AI models (Zeng et al., 2021) to help scale up and optimise the fact-checking workflows. Such models, however, have been trained primarily on datasets of political and other non-COVID-19 misinformation, which has impacted their accuracy in detecting and classifying COVID-19 false claims.

Prior studies of COVID-19 misinformation focused mainly on misinformation detection (Hayawi et al., 2022; Gupta et al., 2021; Hossain et al., 2020), the social engagement with fake news on websites and social platforms (Cui and Lee, 2020), and the ways that misinformation is countered in tweets (Micallef et al., 2020). However, they have largely overlooked the wider online debates about COVID-19 misinformation, such as the conversational threads around false COVID-19 claims and the questions and comments made as part of these. It is absolutely crucial for fact-checkers to have at their disposal models that not only flag misinformation, but can also flag the comments and questions raised in online debates around false claims, so they can address them in debunks.

In particular, this paper aims to address three research questions: **RQ1:** Which social media posts are propagating, questioning or commenting about a false claim? **RQ2:** Does the volume of tweets debunking a misinformation claim correlate with the volume of misinformation tweets? **RQ3:** What are the different kinds of COVID-19 misinformation spreading online? The novel contributions are:

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

[1] The dataset and the annotation codebook are available at https://doi.org/10.5281/zenodo.8131933.

[2] https://www.poynter.org/ifcn/ (Accessed on Feb 1, 2023)

Proceedings of Recent Advances in Natural Language Processing, pages 556–567
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_061

4. A **benchmark experiment** evaluating the performance of misinformation classifiers based on Natural Language Processing (NLP) models on the 8 fine-grained categories.
5. Experimenting with coarse-grained classification which distinguishes (a) **evidence based misinformation classification** from (b) **non-evidence based misinformation classification**. Evidence-based classification aims to classify already verified misinformation given IFCN debunk(s). The harder, non-evidence based task finds social media posts that are likely to be misinformation; however these posts may require human verification.

2 Related Work

2.1 Claim Matching and Automated Fact Checking

There has been rigorous research in the development of automated fact-checking systems (Zeng et al., 2021). As proposed in CLEF CheckThat! Lab task (Nakov et al., 2022, 2021; Barrón-Cedeno et al., 2020), claim matching is one of the pivotal stages to find previously fact-checked claims (Shaar et al., 2020; Vo and Lee, 2020; Singh et al., 2021). The task of claim matching is formulated as an information retrieval task where the false statement from social media is used as a query to a corpus of fact-checked articles. However, in this paper, we do exactly the opposite where we use debunked claims as queries to millions of tweets in order to find relevant tweet matches which include misinformation, debunk, question etc (see Section 3.3 and Section 3.4). We further use this data to train misinformation classifiers on the eight different fine-grained categories (Section 4).

2.2 COVID-19 Datasets

Multiple COVID-19 datasets exist for research purposes, including sentiment analysis of related tweets (Reshi et al., 2022; Nezhad and Deihimi, 2022), and analysis of latent topics and emotions in tweets (Gupta et al., 2021; Almars et al., 2022). Other datasets include COVID-19 scholarly articles (Chen et al., 2020) or provide multilingual Twitter data related to COVID-19 (Gruzd and Mai, 2020).

In terms of datasets that particularly focus on misinformation related to COVID-19, Micallef et al. (2020) investigate the spread of the misinformation and counter-misinformation (debunks) tweets. They present a dataset that focuses on predefined topics and themes (i.e. Fake Cures and 5G Conspiracy Theories), however, the topics of COVID-19 misinformation are fast-evolving. To tackle this, Cui and Lee (2020) present a diverse COVID-19 healthcare misinformation dataset (CoAID) which combines news articles from reliable media outlets to identify instances of misinformation on Twitter. Sharma et al. (2020) label tweets as misinformation if the tweet shares any article or content posted from any of the misinformation sources. However, it is hard to measure the reliability of such data since there is no gold-standard annotation. Hossain et al. (2020) divide COVID-19 misinformation detection into tweet retrieval and stance detection. However, methods evaluated on their dataset are limited to a one-month period. In contrast, our dataset investigates a longer 10-month time span covering tweets from the first and second wave of outbreaks in the US and UK, and relies on professional fact-checkers for debunking evidence.

2.3 COVID-19 Misinformation Detection

Several studies apply rule-based (Singh et al., 2020; Sharma et al., 2020) and machine learning-based methods (Hayawi et al., 2022; Zeng et al., 2021; Micallef et al., 2020) to model the semantic feature in the misinformation. Kou et al. (2022) proposes HC-COVID, a crowdsource knowledge graph based approach to identify and explain misleading COVID-19 claims on social media. Cui and Lee (2020) evaluate the hierarchical attention network (Yang et al., 2016) and its variant dEFEND (Shu et al., 2019) on the CoAID datasets (Cui and Lee, 2020). Meanwhile, Hossain et al. (2020) combine BERTScore (Zhang et al., 2019) with Sentence BERT to identify a tweet's 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, which we investigate in this paper.

Song et al. (2021) propose a classification-aware neural topic model (CANTM) for a COVID-19 disinformation category classification. They also found that the topics of COVID-19 disinformation changed significantly throughout the different stages of the pandemic. Therefore, it is essential to evaluate the performance of disinformation 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 claim out* cross-validation to ensure that there is no topi-

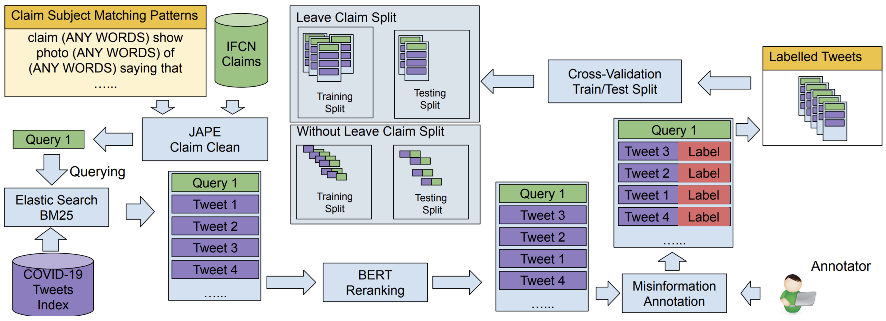


Figure 1: Overall pipeline

cal overlaps between our training and testing data and compare performance against the standard random cross-validation approach (see Section 4.1).

3 Dataset and Annotation

The overall pipeline of dataset annotation is shown in Figure 1. In general, we first collect COVID-19 related tweets based on a set of keywords. Next, we use a subset of fact-checked misinformation claims from the IFCN as queries to retrieve related tweets. The collected tweets are then annotated based on fine-grained categories, and the agreement rates between annotators are evaluated.

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[3] to collect 182,027,646 English tweets spanning 10 months from March to December 2020. Then, we create an 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[4]. 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 the previous research (Song et al., 2021) A pattern matching language – JAPE (Cunningham et al., 2000) 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 two-stage neural ranking (Nogueira and Cho, 2019; Karpukhin et al., 2020), we employ the same for retrieving relevant tweets. In the first retrieval stage, BM25 (Robertson et al., 1995) is utilised to extract the 1,000 most relevant tweets from the Twitter ElasticSearch index. In the second retrieval stage, we employ a pre-trained cross-encoder model[5], which is based on the tiny-BERT architecture (Jiao et al., 2019) and trained on a general information retrieval dataset, specifically the MS MACRO dataset (Nguyen et al., 2016). This model is used to re-rank the retrieved tweets from the first stage 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 specific misinformation in a certain time, since similar misinformation can appear at different stages (e.g.

[3] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data (Accessed on Feb 1, 2023)

[4] https://www.poynter.org/ifcn-covid-19-misinformation/ (Accessed on Feb 1, 2023)

[5] https://huggingface.co/cross-encoder/ms-marco-TinyBERT-L-6 (Accessed on Feb 1, 2023)

Metrics	**Mean Reciprocal Rank**	**Precision@K**				**Mean Average Precision@K**			
K	All	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

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 vaccine roll out).

Table 1 shows the results of our method for retrieving relevant tweet matches. Here, a relevant tweet match can include a tweet which is misinformation, related misinformation, a debunk, a related debunk, a question or comment (please refer to Section 3.4 for the manual annotation process and further details of the classes). We report Mean Reciprocal Rank (MRR), Mean Average Precision (MAP@K) and Precision@K. The results depict high retrieval performance with the MRR of 0.95 and MAP of 0.93 for the top five retrieved tweets. Next, if we consider all the retrieved tweets, we achieve 0.89 MAP, demonstrating the effectiveness of our method for retrieving relevant tweet matches.

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 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-week period. We created groups with three annotators each and we kept the same annotators in each group throughout the 3-week 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 tweet link, 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)

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

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

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 dataset are shown in Table 2 and examples of each class can be found in Appendix A.

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. Bold texts are the coarse-grained labels, and its corresponding fine-grained labels are in the column beneath.

3.5 Data Analysis

This work aims to correlate misinformation and debunk spread with other behaviours (Figure 2). Misinformation tweet volume is notably higher, particularly during the pandemic's start in the first wave in the US and UK. Also, there is a significantly higher volume of 'question and comment' tweets 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 previous findings (Micallef et al., 2020; Mendoza et al., 2010). 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 $\rho = 0.41$, $p < 0.001$). Overall, debunk and misinformation spread rates align, and people comment or question during high misinformation-debunk activity.

Appendix B & C provide detailed analyses of top hashtags (Figure 3) and URL domains (Figure 4) in misinformation and debunk tweets. We observe higher URL frequency in misinformation tweets, potentially including high-credibility sources.

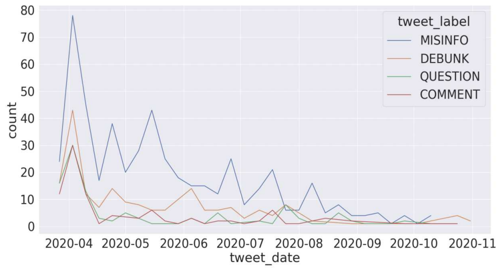


Figure 2: Misinformation, debunk, question and comment tweets volume over time (in weeks).

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. 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.

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 as 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 experiment 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:

- **Standard cross-validation**: 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.
- ***Leave claim out* cross-validation**: 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 realistic testing method to test model performance on 'unseen' misinformation since most of the online misinformation has 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 information (please note, claims are applied in this work as evidence). BERT_Pair and SBERT are evidence dependent models and have been widely applied in Natural Language Inference tasks. The details are as follows:

- **BERT_CLS**: The BERT (Devlin et al., 2018) version used in this experiment is a 24 transformer layer (BERT-large) COVID-Twitter pre-trained (Müller et al., 2020) BERT. Only the parameters in the last transformer encoding layer is unlocked for fine-tuning, the rest of the BERT weights are 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 probability of labels is predicted using a Softmax classifier on the [CLS] representation of the final hidden state.
- **CANTM**: Classification-Aware Neural Topic Model (Song et al., 2021) is a stacked asymmetric variational autoencoder that outputs classification and topic predictions. In this experiment, we only consider the classification output of the CANTM model. The vocabulary size for CANTM is 3,000 with 50 latent topics.
- **Sentence-BERT (SBERT)** (SBERT): We apply SBERT (Reimers and Gurevych, 2019) 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 (Müller et al., 2020) 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 consider-

ation. BERT_Pair is formulated as a pair-wise text classification (Devlin et al., 2018) where the input to the model is [CLS] + Evidence + [SEP] + Tweet_Text + [SEP] and the probability of labels is predicted using a Softmax classifier on the [CLS] representation of the final hidden state. We experiment with two different settings: 1) The results labelled with BERT_Pair_MNLI are trained with the Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018). 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 standard cross-validation setting, 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 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 if the tweet text is misinformation/debunk (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.

In the *leave claim out* cross-validation, all models decreased at least 15% in average F1 measure compared to the standard 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 (see Section 4.1). In other words, models become overfit to the misinformation topics present in the training set. This observation further emphasises the importance of keeping the training data up-to-date 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.

The last row of Table 4 shows the classification performance of the MNLI trained BERT_Pair$_{MNLI}$ model (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, which is an evidence-based task. In the standard 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. Appendix D shows the confusion matrix and a sample of misclassified cases in the fine-grained classification.

In the *leave claim out* cross-validation, all models score average F1 score of less than 0.3, indicating their unreliability 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. No-

	Standard 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
	***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. The highest scores for each metric are in **bold** for both standard and *leave claim out* cross-validation.

	Standard Cross-Validation				***Leave claim out* Cross-Validation**			
	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 class label are R-Debunk:Related Debunk, R-MisInfo:Related Misinformation, COMM:comment, QUES:question, REL:Relevant Other, IR-REL:irrelevant. The highest scores for each metric are in **bold** for both standard and *leave claim out* cross-validation.

tably, the F1 score for the 'Misinformation' class remains consistent with the coarse-grained evidence-based results, likely because it has the highest number of samples in the dataset.

5 Conclusion

This paper presents a fine-grained COVID-19 misinformation dataset, which comprises 1,722 manually annotated tweets across eight categories. Each tweet in the dataset undergoes triple annotation, resulting in a substantial agreement with an averaged Krippendorff's alpha of 0.67. Analysis of the dataset reveals that misinformation tweets have a similar spread rate to debunk tweets. Additionally, we observe that both question and comment tweets have positive correlation with misinformation and debunk tweets. Notably, our findings indicate that misinformation tweets can include URLs from high-credibility sources, shedding light on the potential challenges in identifying misinformation solely based on the source credibility.

Furthermore, the paper presents three misinformation classification benchmark experiments: 1) Non-evidence-based, 2) Evidence-based, and 3) Fine-grained classification. The results of these experiments demonstrate that the baseline models perform well in the standard cross-validation setting across all classification experiments. However, the classification performance dropped significantly in the *leave claim out* cross-validation setting. This emphasises the need for regular updates to the training instances to ensure consistent classification performance over time.

6 Acknowledgement

This research is supported by a UKRI grant EP/W011212/1 and an EU Horizon 2020 grant (agreement no.871042) ("SoBigData++: European Integrated Infrastructure for Social Mining and Big-Data Analytics" (`http://www.sobigdata.eu`)).

7 Ethical Statement and Broader Impact

The experiment processes undertaken has received ethical clearance from the University of Sheffield Ethics Board No. 025371. This research has important implications for countering COVID-19 misinformation on social media by introducing a new dataset for fine-grained classification and informing policy decisions to reduce its negative impact.

References

Abdulqader M Almars, El-Sayed Atlam, Talal H Noor, Ghada ELmarhomy, Rasha Alagamy, and Ibrahim Gad. 2022. Users opinion and emotion understanding in social media regarding covid-19 vaccine. *Computing*, 104(6):1481–1496.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. 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*, pages 215–236. Springer.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

H Cunningham, D Maynard, V Tablan, Hamish Cunningham, H Cunningham, K Bontcheva, W Peters, Y Wilks, Diana Maynard, Hamish Cunningham, et al. 2000. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anatoliy Gruzd and Philip Mai. 2020. COVID-19 Twitter Dataset.

Raj Kumar Gupta, Ajay Vishwanath, and Yinping Yang. 2021. Global reactions to covid-19 on twitter: A labelled dataset with latent topic, sentiment and emotion attributes.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *arXiv preprint arXiv:2011.05773*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, et al. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *ECIR (2)*.

Zahra Bokaee Nezhad and Mohammad Ali Deihimi. 2022. Twitter sentiment analysis from iran about covid 19 vaccine. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 16(1):102367.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Aijaz Ahmad Reshi, Furqan Rustam, Wajdi Aljedaani, Shabana Shafi, Abdulaziz Alhossan, Ziyad Alrabiah, Ajaz Ahmad, Hessa Alsuwailem, Thamer A Almangour, Musaad A Alshammari, et al. 2022. Covid-19 vaccination-related sentiments analysis: a case study using worldwide twitter dataset. In *Healthcare*, volume 10, page 411. MDPI.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. 2020. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv e-prints*, pages arXiv–2003.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.

Iknoor Singh, Kalina Bontcheva, and Carolina Scarton. 2021. The false covid-19 narratives that keep being debunked: A spatiotemporal analysis. *arXiv preprint arXiv:2107.12303*.

Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for covid-19 disinformation categorisation. *PloS one*, 16(2):e0247086.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

WHO. 2020. Novel coronavirus (2019-ncov). `https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf`.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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)*, pages 1112–1122. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*, pages 1480–1489.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Cheng Zhou, Haoxin Xiu, Yuqiu Wang, and Xinyao Yu. 2021. Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on covid-19. *Information Processing & Management*, 58(4):102554.

Appendix

A Dataset Examples

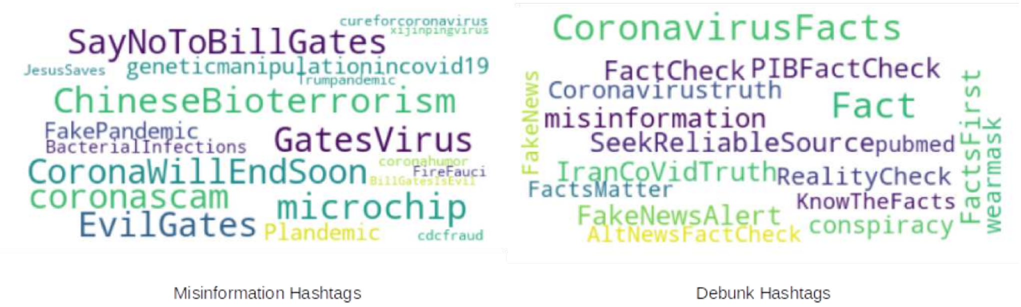
Table 6 shows examples of query and tweets in each class, including misinformation, related misinformation, a debunk, a related debunk, a question, a comment, a relevant and an irrelevant class. Please refer to Section 3.4 in the main paper for details regarding each class.

B Hashtags in Misinformation and Debunk Tweets

Wordclouds of misinformation and debunk tweets is shown in Figure 3. We find that the hashtags are a strong indicator of misinformation as well as debunk tweets. For instance, 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.),

Claim	Tweet	Label
The CDC and other authorities in the US admitted to fake the Covid numbers.	Numbers from #CDC and other agencies are not reported correctly IMO. It is a scare tactic and does not fully allow us to understand #Covid.	Misinformation
More babies die by abortion in two days than all the coronavirus deaths thus far.	There have been approximately 250,000 deaths by abortion in the USA this year so far, approximately 21,000 #coronavirus deaths, yet we are in full #panicmode over #CoronavirusPandemic #wtf #abortion #MSM	Related Misinformation
COVID-19 is a bacterium that is easily treated with aspirin or a coagulant.	Claim- A widely circulated video on social media claims that #Covid19 is a bacteria &amp; which can be treated with aspirin #PIBFactCheck- This is #Fake. Coronavirus is a virus and there is no specific medicinal cure available yet.	Debunk
Steam from boiling oranges kills COVID-19.	#Fact: No scientific evidence to prove that inhaling hot water steam kills #Coronavirus #StayAtHome #GodMorningTuesday #CoronaVirusUpdates #COVID	Related Debunk
Deaths blamed on coronavirus are actually due to the flu.	@TheOfficerTatum @bribohan Wonder if some #Coronavirus "deaths" are actually just FLU or #influenza deaths?	Question
The CDC and other authorities in the US admitted to fake the Covid numbers.	REMINDER: soon the numbers of covid cases in the US will be going through the trump administration and not the CDC. if numbers “start dropping” miraculously take it with a grain of salt.	Comment
COVID-19 cases are “up only because of our big number testing” in the United States.	With the largest number of COVID-19 cases in the world, the United States is seeing disputes heating up over loosening social distancing restrictions and reopening the economy.	Relevant
The novel coronavirus has been artificially created in a laboratory.	Sorrento Therapeutics of San Diego said Friday that an antibody it has been developing proved highly effective in blocking the novel #coronavirus in laboratory experiments — a possible first step in the creation of a drug cocktail to battle COVID-19	Irrelevant

Table 6: Dataset examples



Misinformation Hashtags

Debunk Hashtags

Figure 3: Wordclouds of misinformation and debunk tweets.

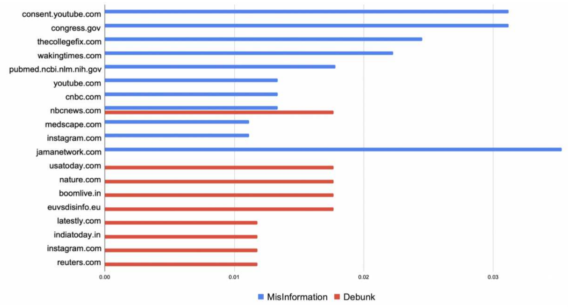
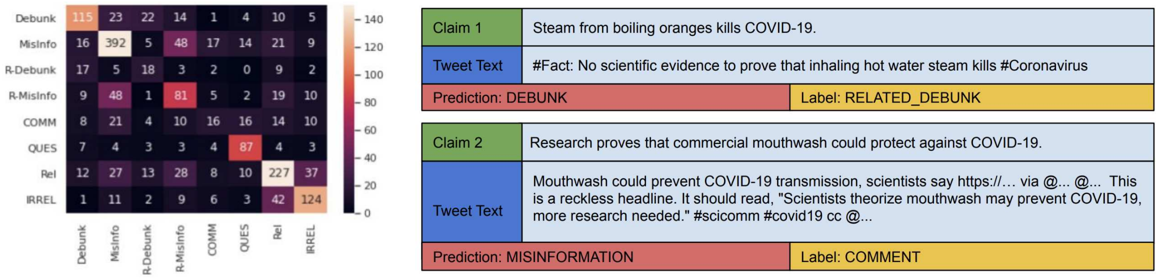


Figure 4: Top 10 frequent URLs found in misinformation and debunk tweets.

and some directly indicate the professional fact-checkers or high-credibility source (e.g., AltNews-FactCheck, pubmed, PIBFactCheck, etc.). Overall, the hashtags in misinformation tweets are found to be more emotional, and debunk hashtags are more related to the professional fact-checkers.

C URL Sources in Misinformation and Debunk Tweets

The top 10 frequent URL domain names found in misinformation and debunk tweets are shown in



	Debunk	MisInfo	R-Debunk	R-MisInfo	COMM	QUES	Rel	IRREL
Debunk	115	23	22	14	1	4	10	5
MisInfo	16	392	5	48	17	14	21	9
R-Debunk	17	5	18	3	2	0	9	2
R-MisInfo	9	48	1	81	5	2	19	10
COMM	8	21	4	10	16	16	14	10
QUES	7	4	3	3	4	87	4	3
Rel	12	27	13	28	8	10	227	37
IRREL	1	11	2	9	6	3	42	124

(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 5: (a) BERT_Pair confusion matrix in the fine-grained classification in standard cross-validation setting. 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) Sample of misclassified cases.

Figure 4. The numbers in horizontal axis are averaged by the 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 a video website (e.g. youtube.com). We also note that URLs in misinformation tweets have high frequency than that of 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.

D BERT_Pair Confusion Matrix

Figure 5 (a) shows the confusion matrix of BERT_Pair results in the fine-grained classification in the standard cross-validation setting. 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 , and the model is unable to catch the difference between them. An example of this error type is presented in Figure 5 (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, 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 5 (b), Claim 2 is an example of comment text. The tweet text quote a misinformation claim 'Mouthwash could prevent COVID-19 transmission' and make a comment that 'more research needed' for this claim.