Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms

Firoj Alam,[1] Fahim Dalvi,[1] Shaden Shaar,[1] Nadir Durrani,[1] Hamdy Mubarak,[1] Alex Nikolov,[2] Giovanni Da San Martino,[3] Ahmed Abdelali,[1] Hassan Sajjad,[1] Kareem Darwish,[1] Preslav Nakov[1]

[1]Qatar Computing Research Institute, HBKU, Qatar
[2]Sofia University "St Kliment Ohridski", Bulgaria
[3]University of Padova, Italy
{fialam,faimaduddin,pnakov}@hbku.edu.qa

Abstract

With the outbreak of the COVID-19 pandemic, people turned to social media to read and to share timely information including statistics, warnings, advice, and inspirational stories. Unfortunately, alongside all this useful information, there was also a new blending of medical and political misinformation and disinformation, which gave rise to the first global infodemic. While fighting this infodemic is typically thought of in terms of factuality, the problem is much broader as malicious content includes not only fake news, rumors, and conspiracy theories, but also promotion of fake cures, panic, racism, xenophobia, and mistrust in the authorities, among others. This is a complex problem that needs a holistic approach combining the perspectives of journalists, fact-checkers, policymakers, government entities, social media platforms, and society as a whole. With this in mind, we define an annotation schema and detailed annotation instructions that reflect these perspectives. We further deploy a multilingual annotation platform, and we issue a *call to arms* to the research community and beyond to join the fight by supporting our crowdsourcing annotation efforts. We perform initial annotations using the annotation schema, and our initial experiments demonstrated sizable improvements over the baselines.

1 Introduction

The year 2020 brought along two remarkable events: the COVID-19 pandemic, and the resulting first global infodemic. The latter thrives in social media, which saw growing use as, due to lockdowns, working from home, and social distancing measures, people spend long hours in social media, where they find and post valuable information, big part of which is about COVID-19. Unfortunately, amidst this rapid influx of information, there was also a spread of disinformation and harmful content in general, fighting which became a matter of utmost importance. In particular, as the COVID-19 outbreak developed into a pandemic, the disinformation about it followed a similar exponential growth trajectory. The extent and the importance of the problem soon lead to international organizations such as the World Health Organization and the United Nations referring to it as the first global *infodemic*. Soon, a number of initiatives were launched to fight this infodemic.

Figure 1: Examples of tweets showing some issues that are important to journalists, fact-checkers, social media platforms, policy makers, government entities, and the society.

The focus of these initiatives was on social media, e.g., building and analyzing large collections of tweet, their content, source, propagators, and spread (Broniatowski et al. 2018; Medford et al. 2020; Mourad et al. 2020; Karami et al. 2021; Leng et al. 2021). Most such efforts were in line with previous work on disinformation detection, which focused almost exclusively on the factuality aspect of the problem, while ignoring the equally important potential to do harm. The COVID-19 infodemic is even more complex, as it goes beyond spreading fake news, rumors, and conspiracy theories, and extends to promote fake cures, panic, racism, xenophobia, and mistrust in the authorities, among others. This is a complex problem that needs a holistic approach combining the perspectives of journalists, fact-checkers, policymakers, government entities, social media platforms, and society.

Here we define a comprehensive annotation schema that goes beyond factuality and potential to do harm, extending to information that could be potentially useful, e.g., for government entities to notice or for social media to promote. Information about a possible cure for COVID-19 should get the attention of a fact-checker, and if proven false, as in the example in Figure 1(a), it should be flagged with a warning or even removed from the social media platform to prevent its further spread; it might also need a response by a public health official. However, if proven truthful, it might instead be promoted in view of the high public interest in the matter. Our annotation schema further covers some categories of good posts, including giving advice, asking a question, discussing action taken, possible cure as in Figure 1(b), or calling for action as in Figure 1(c). Such posts could be useful for journalists, policymakers, and society as a whole.

We organize the annotations around seven questions, asking whether a tweet (1) contains a verifiable factual claim, (2) is likely to contain false information, (3) is of interest to the general public, (4) is potentially harmful to a person, a company, a product, or society, (5) requires verification by a fact-checker, (6) poses harm to society, or (7) requires the attention of a government entity. Annotating so many aspects is challenging and time-consuming. Moreover, the answer to some of the questions is subjective, which means we really need multiple annotators per example, as we have found in our preliminary manual annotations. Our contributions can be summarized as follows:

- We develop comprehensive guidelines that combine the perspectives and the interests of journalists, fact-checkers, social media platforms, policymakers, and the society as a whole.
- We develop a volunteer-based crowd annotation platform based on Micromappers[1], and we invite volunteers to join our annotation efforts.
- We annotate initial datasets covering English and Arabic.
- We demonstrate sizable improvements over the baselines when using both coarse and fine-grained labels.

2 Call to Arms

We invite volunteers to join our crowdsourcing annotation efforts and to label some new tweets, thus supporting the fight against the COVID-19 infodemic. We make all resulting annotations publicly available.[2] As of present, we focus on English and Arabic; however, we plan to add more languages in the future.

3 Related Work

"Fake News", Disinformation, and Misinformation: There has been a lot of interest in recent years in identifying disinformation, misinformation, and "fake news", which thrive in social media. The studies of (Lazer et al. 2018) and (Vosoughi, Roy, and Aral 2018) in *Science* offered a general overview and discussion on the science of "fake news" and of the process of proliferation of true and false news online. There have also been several interesting surveys. Shu et al. (2017) studied information dissemination and consumption in social media. Thorne and Vlachos (2018) took a fact-checking perspective on "fake news" and related problems. Li et al. (2016) covered truth discovery in general. Some very recent surveys focused on stance for misinformation and disinformation detection (Hardalov et al. 2021), on automatic fact-checking to assist human fact-checkers (Nakov et al. 2021a), on predicting the factuality and the bias of entire news outlets (Nakov et al. 2021d), on multimodal disinformation detection (Alam et al. 2021), and on abusive language in social media (Nakov et al. 2021c).

Fact-Checking: Research in this direction includes *fact-checking*, i.e., verifying the veracity of the claim in textual and imagery content, and *check-worthiness*, deciding whether a claim is worthy of investigation by a professional fact-checker. There have been a number of professional organizations working on *fact-checking*,[3] and a number of dataset have been developed by the NLP research community to develop models for automatic fact-checking. Some of the larger datasets include the *Liar, Liar* dataset of 12.8K claims from PolitiFact (Wang 2017), *ClaimsKG* dataset and system (Tchechmedjiev et al. 2019) of 28K claims from 8 fact-checking organizations, the *MultiFC* dataset of 38K claims from 26 fact-checking organizations (Augenstein et al. 2019), and the 10K claims *Truth of Various Shades* (Rashkin et al. 2017) dataset, among other smaller-size ones. A number of datasets have also been developed as part of shared tasks. In most cases, they did not rely on fact-checking websites, but performed their own annotation, either (a) manually, e.g., the SemEval-2017 task 8 (Derczynski et al. 2017) and the SemEval-2019 task 7 (Gorrell et al. 2019) on Determining Rumour Veracity and Support for Rumours (RumourEval), the SemEval-2019 task 8 on Fact-Checking in Community Question Answering Forums (Mihaylova et al. 2019), the CLEF 2018–2021 CheckThat! Lab (Nakov et al. 2018; Elsayed et al. 2019; Barrón-Cedeño et al. 2020; Nakov et al. 2021b), which featured both English and Arabic, or (b) using crowdsourcing, e.g., the FEVER 2018–2019 tasks on Fact Extraction and VERification, which focused on fact-checking made-up claims about content present in Wikipedia (Thorne et al. 2019). Some non-English datasets have also been developed, e.g., Baly et al. (2018) developed a dataset of 402 Arabic claims extracted from Verify-SY.

Check-Worthiness: As the detected claims can be large in volume, it is important to identify which claims are check-worthy. A manually labeled dataset for check-worthiness was used in the ClaimBuster system (Hassan, Li, and Tremayne 2015). Gencheva et al. (2017) developed a dataset of political debates with labels collected from fact-checking websites. This dataset was used in the ClaimRank system (Jaradat et al. 2018), and it was extended (Vasileva et al. 2019) and used in the CLEF CheckThat! labs 2018-2021 (Nakov et al. 2018; Elsayed et al. 2019; Barrón-Cedeño et al. 2020; Nakov et al. 2021b).

Fighting the COVID-19 Infodemic: There have been a

[1] http://micromappers.qcri.org

[2] Our data: http://doi.org/10.7910/DVN/XYK2UE

[3] https://en.wikipedia.org/wiki/List_of_fact-checking_websites

number of COVID-19 Twitter datasets: many without labels, other using distant supervision, and very few manually annotated. Some large datasets include a multi-lingual dataset of 123M tweets (Chen, Lerman, and Ferrara 2020), another one of 152M tweets (Banda et al. 2020), a billion-scale dataset of 65 languages and 32M geo-tagged tweets (Abdul-Mageed et al. 2021), and the GeoCoV19 dataset, consisting of 524M multilingual tweets, including 491M with GPS coordinates (Qazi, Imran, and Ofli 2020). There have also two Arabic datasets, some without manual annotations (Alqurashi, Alhindi, and Alanazi 2020), and some with (Haouari et al. 2021; Mubarak and Hassan 2021). Cinelli et al. (2020) studied rumor amplification in five social media platforms, where rumors were labeled using distant supervision. In contrast, we have careful manual annotation and multiple labels. Zhou et al. (2020) created the ReCOVery dataset, which combines news articles about COVID-19 with tweets. Vidgen et al. (2020) studied COVID-19 prejudices using a manually labeled dataset of 20K tweets with the following labels: hostile, criticism, prejudice, and neutral. The closest work to ours is that of Song et al. (2021), who collected a dataset of false and misleading claims about COVID-19 from IFCN Poynter, which they manually annotated with ten disinformation categories: (1) Public authority, (2) Community spread and impact, (3) Medical advice, self-treatments, and virus effects, (4) Prominent actors, (5) Conspiracies, (6) Virus transmission, (7) Virus origins and properties, (8) Public reaction, and (9) Vaccines, medical treatments, and tests, and (10) Cannot determine. Their categories partially overlap with ours, but ours are broader and account for more perspectives. Moreover, we cover both true and false claims, we focus on tweets (while they have general claims), and we cover both English and Arabic (they only cover English). Other related work is FakeCovid (Shahi and Nandini 2020), a multilingual cross-domain dataset consisting of manually labeled 1,951 articles. The study by (Pulido et al. 2020) analyzed 1,000 tweets and categorized them based on factuality: *(i)* False information, *(ii)* Science-based evidence, *(iii)* Fact-checking tweets, *(iv)* Mixed information, *(v)* Facts, *(vi)* Facts, *(vii)* Other, and *(viii)* Not valid. Finally, Ding et al. (2020) have a position paper discussing the challenges in combating the COVID-19 infodemic in terms of data, tools, and ethics. See also a recent survey by Shuja et al. (2020).

4 Annotation Setup

Below, we present the annotation schema that we developed after a lot of analysis and discussion, and which we refined during the pilot annotations. We then present the annotation platform and interface we used.

4.1 Annotation Schema and Instructions

We designed the annotation instructions after careful analysis and discussion, followed by iterative refinement based on observations from the pilot annotation. Our annotation schema is organized into seven questions.

Q1: Does the tweet contain a verifiable factual claim? This is an objective question, and it proved very easy to annotate. Positive examples include[4] tweets that state a definition, mention a quantity in the present or in the past, make a verifiable prediction about the future, reference laws, procedures, and rules of operation, discuss images or videos, and state correlation or causation, among others.

We show the annotator the tweet text only. This is a *Yes*/*No* question, but we also have a *Don't know or can't judge* answer. If the annotator selects *Yes*, then questions 2–5 are to be answered as well; otherwise, they are skipped automatically (see Section 4.2).

Q2: To what extent does the tweet appear to contain false information? This question asks for a subjective judgment; it does not ask for annotating the actual factuality of the claim in the tweet, but rather whether the claim *appears* to be false. For this question (and for all subsequent questions), we show the tweet as it is displayed in the Twitter feed, which can reveal some useful additional information, e.g., a link to an article from a reputable information source could make the annotator more likely to believe that the claim is true. The annotation is on a 5-point ordinal scale:

1. *NO, definitely contains no false information*
2. *NO, probably contains no false information*
3. *not sure*
4. *YES, probably contains false information*
5. *YES, definitely contains false information*

Q3: Will the tweet have an effect on or be of interest to the general public? Generally, claims that contain information related to potential cures, updates on number of cases, on measures taken by governments, or discussing rumors and spreading conspiracy theories should be of general public interest. Similarly to Q2, the labels are defined on a 5-point ordinal scale; however, unlike Q2, this question is partially objective (the *YES*/*NO* part) and partially subjective (the *definitely*/*probably* distinction).

1. *NO, definitely not of interest*
2. *NO, probably not of interest*
3. *not sure*
4. *YES, probably of interest*
5. *YES, definitely of interest*

Q4: To what extent is the tweet harmful to the society, person(s), company(s) or product(s)? This question asks to identify tweets that can negatively affect society as a whole, but also specific person(s), company(s), product(s). The labels are again on a 5-point ordinal scale, and, similarly to Q3, this question is partially objective (*YES*/*NO*) and partially subjective (*definitely*/*probably*).

1. *NO, definitely not harmful*
2. *NO, probably not harmful*
3. *not sure*
4. *YES, probably harmful*
5. *YES, definitely harmful*

[4]This is influenced by (Konstantinovskiy et al. 2018).

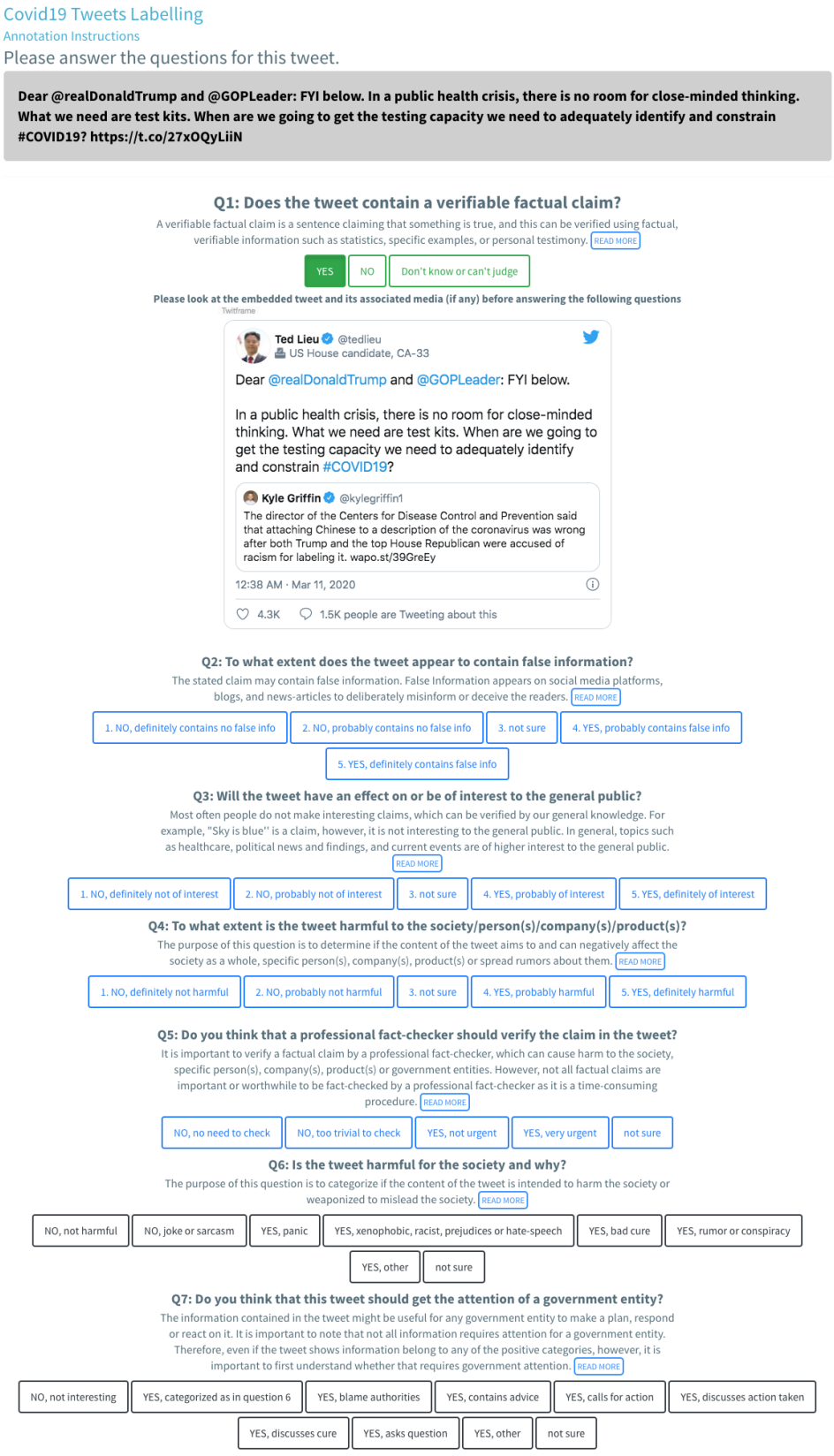

Figure 2: The platform for an English tweet: a Yes answer for Q1 has shown questions Q2–Q7 and their answers.

Figure 3: The platform for an Arabic tweet: a No answer for Q1 means that only Q6 and Q7 would be shown. (English translation of the Arabic text in the tweet: We must prevent the collapse of the healthcare system. The Ministry of Public Health will cure the infected people, but the spread of the infection puts the elderly and our beloved ones in danger. That is why we say #StayHomeForQatar, and we will succeed...)

Q5: Do you think that a professional fact-checker should verify the claim in the tweet? This question asks for a subjective opinion. Yet, its answer should be informed by the answer to questions Q2, Q3 and Q4, as a check-worthy factual claim is probably one that is likely to be false, is of public interest, and/or appears to be harmful. Here the answers are not on an ordinal scale, but rather focus on the reason why there is or is not a need to fact-check the tweet:

A. *NO, no need to check*: there is no need to fact-check the claims(s) made in the tweet, e.g., because they are not interesting, make a joke, etc.

B. *NO, too trivial to check*: the tweet is worth fact-checking, but this does not require a professional fact-checker, i.e., a non-expert might be able to fact-check it easily, e.g., by using reliable sources such as the official website of the World Health Organization, Wikipedia, etc.

C. *YES, not urgent*: the tweet should be fact-checked by a fact-checker, but this is not urgent, nor is it critical.

D. *YES, very urgent*: the tweet can cause immediate harm to a large number of people, and thus it should be fact-checked as soon as possible by a professional fact-checker.

E. *not sure*: the tweet does not contain enough information to allow for a clear judgment on whether it is worth fact-checking, or the annotator is simply not sure.

Q6: Is the tweet harmful to the society and why? This is an objective question. It asks whether the tweet is harmful to the society (unlike Q4, which covers broader harm, e.g., to persons, companies, and products). It further asks to categorize the nature of the harm, if any. Similarly to Q5 (and unlike Q4), the answers are categorical and are not on an ordinal scale.

A. *NO, not harmful*: the tweet is not harmful to the society.

B. *NO, joke or sarcasm*: the tweet contains a joke or expresses sarcasm.

C. *not sure*: the content of the tweet makes it hard to make a judgment.

D. *YES, panic*: the tweet can cause panic, fear, or anxiety.

E. *YES, xenophobic, racist, prejudices, or hate-speech*: the tweet contains a statement that relates to xenophobia, racism, prejudices, or hate speech.

F. *YES, bad cure*: the tweet promotes a questionable cure, medicine, vaccine, or prevention procedures.

G. *YES, rumor, or conspiracy*: the tweet spreads rumors or conspiracy theories.

H. *YES, other*: the tweet is harmful, but it does not belong to any of the above categories.

Q7: Do you think that this tweet should get the attention of a government entity? This question asks for a subjective judgment (unlike Q6 which was objective) about whether the target tweet should get the attention of a government entity or of policy makers in general. Similarly to Q5 and Q6, the answers to this question are categorical and are not on an ordinal scale.

A. *NO, not interesting*: the tweet is not interesting for any government entity.

B. *not sure*: the content of the tweet makes it hard to make a judgment.

C. *YES, categorized as in Q6*: a government entity should pay attention to this tweet as it was labeled with some of the *YES* sub-categories in Q6.

D. *YES, other*: the tweet needs the attention of a government entity, but it cannot be labeled as any of the above categories.

E. *YES, blames authorities*: the tweet blames government authorities or top politicians.

F. *YES, contains advice*: the tweet contains advice about some COVID-19 related social, political, national, or international issues that might be of interest to a government entity.

G. *YES, calls for action*: the tweet states that some government entities should take action on a particular issue.

H. *YES, discusses action taken*: the tweet discusses specific actions or measures taken by governments, companies, or individuals regarding COVID-19.

I. *YES, discusses cure*: the tweet discusses possible cure, vaccine or treatment for COVID-19.

J. *YES, asks a question*: the tweet raises question that might need an official answer.

A notable property of our schema is that the fine-grained labels can be easily transformed into coarse-grained binary YES/NO labels, i.e., all no* labels could be merged into a *NO* label, and all yes* labels can become *YES*. Note also that some questions (i.e., Q2, Q3, and Q4) are on an ordinal scale, and thus can be addressed using ordinal regression.

Finally, note that even though our annotation instructions were developed to analyze the COVID-19 infodemic, they can be potentially adapted for other kinds of global crises, where taking multiple perspectives into account is desirable.

4.2 Annotation Platform

Our crowd-sourcing annotation platform is based on MicroMappers,[1] a framework that was used for several disaster-related social media volunteer annotation campaigns in the past. We configured MicroMappers to allow labeling COVID-19 tweets in English and Arabic for all seven questions. Initially, the interface only shows the text of the tweet and the answer options for Q1. Then, depending on the selected answer, it dynamically shows either Q2-Q7 or Q6-Q7. After Q1 has been answered, it shows not just the text of the tweet, but its actual look and feel as it appears on Twitter. The annotation instructions are quickly accessible at any moment for the annotators to check.

Figure 2 shows an example of an English tweet, where the answer *Yes* was selected for Q1, which has resulted in displaying the tweet as it would appear in Twitter as well as showing all the remaining questions with their associated answers. Figure 3 shows an Arabic example, where a *No* answer was selected,[5] which has resulted in showing questions Q6 and Q7 only. The use of annotation platform has reduced our in-house annotation efforts significantly, cutting the annotation time by half compared to using a spreadsheet, and we expect similar time savings for general crowdsourcing annotations. The platform is collaborative in nature, and multiple annotators can work on it simultaneously. In order to ensure the quality of the annotations, we have configured the platform to require five annotators per tweet.

5 Pilot Annotation Dataset

5.1 Data for the Pilot Annotation

We collected frequent tweets (with at least 500 retweets) about COVID-19 in March 2020, in English and Arabic.

5.2 Annotation

We performed a pilot annotation in order to test the platform and to refine the annotation guidelines. We annotated 504 English tweets for questions Q1, Q6, and Q7; however, we have 305 tweets for questions Q2, Q3, Q4, and Q5 as they are only annotated if the answer to Q1 is *Yes*. Similarly, for Arabic, we have 218 tweets for Q1, Q6, and Q7, and 140 tweets for Q2, Q3, Q4, and Q5.

We performed the annotation in three stages. In the first stage, 2–5 annotators independently annotated a batch of 25-50 examples. In the second stage, these annotators met to discuss and to try to resolve the cases of disagreement. In the third stage, any unresolved cases were discussed in a meeting involving all authors of this paper.

In stages two and three, we further discussed whether handling the problematic tweets required adjustments or clarifications in the annotation guidelines. In case of any such change for some questions, we reconsidered all previous annotations for that question in order to make sure the annotations reflected the latest version of the annotation guidelines.

5.3 Annotation Agreement

In the process of annotation, we were calculating the current inter-annotator agreement. Fleiss Kappa was generally high for objective questions, e.g., it was over 0.9 for Q1, and around 0.5 for Q6. For subjective and partially subjective questions, the scores ranged around 0.4 and 0.5, with the notable exception of Q5 with 0.8. Note that values of Kappa of 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.0 correspond to fair, moderate, substantial and perfect agreement, respectively (Landis and Koch 1977).

[5]Note that this answer is actually wrong, as there are verifiable factual claims in the tweet. Here, it was selected for demonstration purposes only.

Exp.	Class labels	EN	AR
Q1: Does the tweet contain a verifiable factual claim?		**504**	**218**
Bin	No	199	78
	Yes	305	140
Q2: To what extent does the tweet appear to contain false information?		**305**	**140**
Multi	No, definitely contains no false info	46	31
	No, probably contains no false info	177	62
	not sure	45	5
	Yes, probably contains false info	25	40
	Yes, definitely contains false info	12	2
Bin	No	223	93
	Yes	37	42
Q3: Will the tweet's claim have an effect on or be of interest to the general public?		**305**	**140**
Multi	No, definitely not of interest	10	1
	No, probably not of interest	46	5
	not sure	8	9
	Yes, probably of interest	180	76
	Yes, definitely of interest	61	49
Bin	No	56	6
	Yes	241	125
Q4: To what extent does the tweet appear to be harmful to society, person(s), company(s) or product(s)?		**305**	**140**
Multi	No, definitely not harmful	111	68
	No, probably not harmful	67	21
	not, sure	2	3
	Yes, probably harmful	67	46
	Yes, definitely harmful	58	2
Bin	No	178	89
	Yes	125	48
Q5: Do you think that a professional fact-checker should verify the claim in the tweet?		**305**	**140**
Multi	No, no need to check	81	22
	No, too trivial to check	64	55
	Yes, not urgent	117	48
	Yes, very urgent	43	15
Bin	No	145	77
	Yes	160	63

Table 1: Distribution for the English and the Arabic datasets for questions 1 to 5. In the rows with a question, we show the total number of tweets for the respective language. For the binary task (Bin), we map all multiclass (Multi) Yes* labels to Yes, and the No* labels to No, and we further drop the not sure labels.

5.4 Data Statistics

Table 1 and 2 shows statistics about the annotations. While we focus the following analysis on English tweets, the distribution of the Arabic ones is similar.

The class distribution for Q1 is quite balanced, (61% YES and 39% NO examples). Recall that only the tweets that are labeled as factual were annotated for Q2-5. For question Q2, the label "No, probably contains no false info" is frequent, which means that most tweets considered credible.

Exp.	Class labels	EN	AR
Q6: Is the tweet harmful for society and why?		**504**	**218**
Multi	No, joke or sarcasm	62	2
	No, not harmful	333	159
	not sure	2	0
	Yes, bad cure	3	1
	Yes, other	25	5
	Yes, panic	23	12
	Yes, rumor conspiracy	42	33
	Yes, xenophobic racist prejudices or hate speech	14	6
Bin	No	395	161
	Yes	107	57
Q7: Do you think that this tweet should get the attention of any government entity?		**504**	**218**
Multi	No, not interesting	319	163
	not sure	6	0
	Yes, asks question	2	0
	Yes, blame authorities	81	13
	Yes, calls for action	8	1
	Yes, classified as in question 6	34	30
	Yes, contains advice	9	1
	Yes, discusses action taken	12	6
	Yes, discusses cure	5	4
	Yes, other	28	0
Bin	No	319	163
	Yes	179	55

Table 2: Distribution for the English and the Arabic datasets for question 6 and 7.

Out of 305 tweets labeled for Q2, about 73% are judged to contain no false information, whereas 12% were categorized as "not sure", and 15% as "contains false information", either "probably" or "definitely".

For Q3, which asks whether *the tweet is of interest to the general public*, the distribution is skewed towards *Yes* in 79% of the cases. This can be attributed to the tweets having been selected based on frequency of retweets and likes.

For Q4, which asks whether *the tweet is harmful to the society*, the labels for the tweets vary from not harmful to harmful, covering all cases, without huge spikes.

For Q5, which asks whether *a professional fact-checkers should verify the claim*, the majority of the cases were labeled as either "Yes, not urgent" (38%) or "No, no need to check" (27%). It appears that a professional fact-checker should verify the claims made in the tweets immediately in only a small number of cases (14%).

For questions Q3-5, the "not sure" cases were very few. Yet, such cases were substantially more prevalent in the case of Q2. Identifying potentially false claims (Q2) is challenging, as it might require external information. When annotating Q2, the annotators were shown the entire tweet, and they could further open tweet and see the entire thread in Twitter.

For Q6, most of the tweets were classified as "not harmful" for the society or as a "joke or sarcasm". From the critical classes, 3% of the tweets are classified as containing "xenophobic, racist, prejudices or hate speech", and 5% as "spreading panic". For Q7, it is clear that, in the majority of cases (64%), the tweets are not of interest to policy makers; yet, 16% of the tweets blame the authorities.

6 Experiments and Evaluation

6.1 Experimental Setup

We performed experiments using both binary and multiclass settings. We first performed standard pre-processing of the tweets: removing hash tags and other symbols, and replacing URLs and usernames by special tags. Due to the small size of the datasets, we used 10-fold cross validation. To tune the hyper-parameters of the models, we split each training fold into $\texttt{train}_{train}$ and $\texttt{train}_{dev}$ parts, and we used the latter for finding the best hyper-parameter values.

Models Large-scale pre-trained transformers have achieved state-of-the-art performance for several NLP tasks. We experimented with such models and binary vs. multiclass, low-resource task scenarios. More specifically, we used BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and ALBERT (Lan et al. 2019) for English, and multilingual BERT (mBERT), XLM-r (Conneau et al. 2020) and AraBERT (Baly, Hajj et al. 2020) for Arabic. In addition to pre-trained models, we also evaluated the performance of static-embedding based classification using FastText (Joulin et al. 2017).

For transformer-based models, we fine-tuned each model using the default settings for three epochs as described in (Devlin et al. 2019). Due to instability, we performed ten runs of each experiment using different random seeds, and we picked the model that performs the best on the development set. For FastText, we used embeddings trained on Common Crawl.

Evaluation Measure We report weighted F1 score, as it takes class imbalance into account.

6.2 Results

Baseline We use a simple majority class baseline. Note that for questions with highly imbalanced label distribution, it can achieve very high scores. For example, for Q3 in the binary setting for Arabic, 125 out of 131 tweets are in the 'Yes' category, which yields a balanced F1 score of 93% for the majority class baseline.

Binary Classification The first part of Table 3 presents the results for binary classification using various models.

Results for English: We can see that all models performed better than the majority class baseline and FastText, confirming the efficacy of transformers. Comparing various pre-trained models, we can see that BERT outperformed all other models on six out of the seven tasks, while ALBERT performed the worst in most of the cases. For Q1, RoBERTa and mBERT performed better than BERT, with RoBERTa performing the best.

Results for Arabic: We can see that for all tasks except for Q3 in Arabic (which has very skewed distribution), the models performed better than the majority class baseline. Unlike English, this time, there was no model that outperformed the rest overall. We can see that XLM-r performed worse, that mBERT outperformed all the other models for four out of seven tasks, and that AraBERT performed better than other models for Q4.

Interestingly, FastText performed very well on many tasks, achieving the best overall results on Q5. This could be due to it using character n-grams, which can be important for a morphologically rich language such as Arabic.

Multiclass Classification The second part of Table 3 shows the results in the multiclass setting. The *Cls* column shows the number of classes per task, and we can see that the number of classes now increases from 2 to 5–10, depending on the question. This makes the classification tasks much harder, which is reflected in the substantially lower weighted F1 scores, both for the baselines and for the models we experimented with.

Results for English: We can see that all models performed better than the majority class. The most successful one was mBERT, which performed the best in four out of six tasks. Interestingly, mBERT outperformed BERT in several cases.

Results for Arabic: This time, FastText outperformed all transformers models. Once again, this can be due to it using embeddings for character n-grams, which makes it more robust to morhphological variations in the input, including possible typos. This could also indicate the training data not being sufficient to optimize the large number of parameters in the transformer models.

7 Conclusion and Future Work

In a bid to effectively counter the first global infodemic related to COVID-19, we have argued for the need for a holistic approach combining the perspectives of journalists, fact-checkers, policymakers, government entities, social media platforms, and society. With this in mind and in order to reduce the annotation effort and to increase the quality of the annotations, we have developed a volunteer-based crowd annotation tools based on the MicroMappers platform. Now, we issue a *call to arms* to the research community and beyond to join the fight by supporting our crowd-sourcing annotation efforts. We plan to support the annotation platforms with fresh tweets. We further plan to release annotation platforms for other languages. Finally, we plan regular releases of the data obtained thanks to the crowdsourcing efforts.

Acknowledgments

This research is part of the Tanbih project,[6] developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news", propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

Broader Impact

Our dataset was collected from Twitter, following its terms of service. It can enable analysis of social media content, which could be of interest to practitioners, professional fact-checker, journalists, social media platforms, and policy makers. Our models can help fight the infodemic, and they could support analysis and decision making for the public good. However, they could also be misused by malicious actors.

[6] http://tanbih.qcri.org

		English						Arabic				
Q.	**Cls**	**Maj.**	**FastText**	**BERT**	**mBERT**	**RoBERTa**	**ALBERT**	**Maj.**	**FastText**	**mBERT**	**AraBERT**	**XLM-r**
Binary (Coarse-grained)												
Q1	2	45.6	**72.8**	**87.6**	**88.3**	**90.6**	**86.5**	50.2	**75.8**	**88.1**	**82.6**	**76.9**
Q2	2	79.2	**82.6**	**86.9**	**83.1**	**82.9**	**83.9**	56.2	**68.2**	**79.1**	**71.1**	**60.2**
Q3	2	72.7	**77.2**	**84.3**	**81.6**	**80.8**	**79.6**	93.2	93.2	89.2	77.8	89.2
Q4	2	43.5	**69.6**	**84.0**	**82.7**	**83.8**	**78.5**	51.2	**79.2**	**78.5**	**80.4**	**69.0**
Q5	2	36.1	**63.1**	**81.3**	**80.0**	**73.7**	**72.7**	39.0	**78.6**	**76.4**	**76.1**	**66.5**
Q6	2	69.3	**71.6**	**86.1**	**76.8**	**81.0**	**79.2**	62.7	**79.4**	**80.4**	**77.3**	**64.6**
Q7	2	50.0	**69.9**	**89.3**	**81.9**	**84.7**	**79.0**	64.0	**74.1**	**78.5**	**77.9**	64.0
Multiclass (Fine-grained)												
Q2	5	42.6	**44.0**	**48.5**	**52.2**	**46.6**	**44.8**	27.2	**47.4**	**42.8**	**42.1**	**37.4**
Q3	5	43.8	**48.3**	**57.6**	**45.1**	**50.9**	**45.4**	**38.2**	**83.1**	27.0	21.4	20.0
Q4	5	19.4	**35.5**	**41.6**	**42.9**	**44.1**	**39.5**	31.8	**54.4**	**43.7**	**44.9**	**34.2**
Q5	5	21.3	**37.6**	**50.4**	**52.3**	**50.3**	**48.0**	22.2	**77.2**	**59.0**	**57.7**	**46.1**
Q6	8*	52.6	**53.9**	**57.2**	**62.7**	**58.4**	**56.5**	61.5	**79.3**	40.9	38.9	44.5
Q7	10*	49.1	**57.8**	**54.6**	**58.7**	**55.2**	**53.5**	64.0	**75.7**	**66.3**	63.9	64.0

Table 3: Experiments using different models. Binary and multiclass results (weighted F1), for English and Arabic, using various Transformers and FastText. The results that improve over the majority class baseline (Maj.) are in bold, and the best system is underlined. Legend: Q. – question, Cls – number of classes, the * in Q6 and Q7 is a reminder that for Arabic there are 7 classes (not 8 and 10 as for English).

References

Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E. M. B.; Pabbi, D.; Verma, K.; and Lin, R. 2021. Mega-COV: A Billion-Scale Dataset of 100+ Languages for COVID-19. In *EACL*.

Alam, F.; Cresci, S.; Chakraborty, T.; Silvestri, F.; Dimitrov, D.; Martino, G. D. S.; Shaar, S.; Firooz, H.; and Nakov, P. 2021. A Survey on Multimodal Disinformation Detection. *arXiv/2103.12541* .

Alqurashi, S.; Alhindi, A.; and Alanazi, E. 2020. Large Arabic Twitter Dataset on COVID-19. *arXiv/2004.04315* .

Augenstein, I.; Lioma, C.; Wang, D.; Chaves Lima, L.; Hansen, C.; Hansen, C.; and Simonsen, J. G. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *EMNLP-IJCNLP*, 4685–4697.

Baly, F.; Hajj, H.; et al. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Workshop on OSACT, LREC*, 9–15.

Baly, R.; Mohtarami, M.; Glass, J.; Màrquez, L.; Moschitti, A.; and Nakov, P. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *NAACL*, 21–27.

Banda, J. M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; and Chowell, G. 2020. A large-scale COVID-19 Twitter Chatter Dataset for Open Scientific Research – An International Collaboration. *arXiv:2004.03688* .

Barrón-Cedeño, A.; Elsayed, T.; Nakov, P.; Martino, G. D. S.; Hasanain, M.; Suwaileh, R.; and Haouari, F. 2020. CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media. In *ECIR*, 499–507.

Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *APHA* 108(10): 1378–1384.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR* 6(2): e19273.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 Social Media Infodemic. *Sci. Reports* 10(1): 1–10.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*, 8440–8451.

Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Zubiaga, A. 2017. SemEval-2017 Task 8: RumourEval: Determining Rumour Veracity and Support for Rumours. In *SemEval*, 60–67.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Ding, K.; Shu, K.; Li, Y.; Bhattacharjee, A.; and Liu, H. 2020. Challenges in Combating COVID-19 Infodemic – Data, Tools, and Ethics. *arXiv:2005.13691* .

Elsayed, T.; Nakov, P.; Barrón-Cedeño, A.; Hasanain, M.; Suwaileh, R.; Da San Martino, G.; and Atanasova, P. 2019. CheckThat! at CLEF 2019: Automatic Identification and Verification of Claims. ECIR, 309–315.

Gencheva, P.; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; and Koychev, I. 2017. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In *RANLP*, 267–276.

Gorrell, G.; Aker, A.; Bontcheva, K.; Derczynski, L.; Kochkina, E.; Liakata, M.; and Zubiaga, A. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *SemEval*, 845–854.

Haouari, F.; Hasanain, M.; Suwaileh, R.; and Elsayed, T. 2021. ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. In *Arabic NLP workshop*.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2021. A Survey on Stance Detection for Mis- and Disinformation Identification. *arXiv/2103.00242* .

Hassan, N.; Li, C.; and Tremayne, M. 2015. Detecting Check-Worthy Factual Claims in Presidential Debates. In *CIKM*, 1835–1838.

Jaradat, I.; Gencheva, P.; Barrón-Cedeño, A.; Màrquez, L.; and Nakov, P. 2018. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. In *NAACL-HLT*, 26–30.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*, 427–431.

Karami, A.; Lundy, M.; Webb, F.; Turner-McGrievy, G.; McKeever, B. W.; and McKeever, R. 2021. Identifying and Analyzing Health-Related Themes in Disinformation Shared by Conservative and Liberal Russian Trolls on Twitter. *Int. J. Environ. Res. Public Health* 18(4): 2159.

Konstantinovskiy, L.; Price, O.; Babakar, M.; and Zubiaga, A. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *arXiv:1809.08193* .

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942* .

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *biometrics* 159–174.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The Science of Fake News. *Science* 359(6380): 1094–1096.

Leng, Y.; Zhai, Y.; Sun, S.; Wu, Y.; Selzer, J.; Strover, S.; Zhang, H.; Chen, A.; and Ding, Y. 2021. Misinformation During the COVID-19 Outbreak in China: Cultural, Social and Political Entanglements. *IEEE Trans. on Big Data* 7(1): 69–80.

Li, Y.; Gao, J.; Meng, C.; Li, Q.; Su, L.; Zhao, B.; Fan, W.; and Han, J. 2016. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.* 17(2): 1–16.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692* .

Medford, R. J.; Saleh, S. N.; Sumarsono, A.; Perl, T. M.; and Lehmann, C. U. 2020. An "Infodemic": Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak. *OFID* 7(7).

Mihaylova, T.; Karadzhov, G.; Atanasova, P.; Baly, R.; Mohtarami, M.; and Nakov, P. 2019. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In *SemEval*, 860–869.

Mourad, A.; Srour, A.; Harmanai, H.; Jenainati, C.; and Arafeh, M. 2020. Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions. *IEEE TNSM* 17(4): 2145–2155.

Mubarak, H.; and Hassan, S. 2021. ArCorona: Analyzing Arabic Tweets in the Early Days of Coronavirus (COVID-19) Pandemic. *arXiv:2012.01462* .

Nakov, P.; Barrón-Cedeño, A.; Elsayed, T.; Suwaileh, R.; Màrquez, L.; Zaghouani, W.; Atanasova, P.; Kyuchukov, S.; and Da San Martino, G. 2018. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In *CLEF*, 372–387.

Nakov, P.; Corney, D.; Hasanain, M.; Alam, F.; Elsayed, T.; Barrón-Cedeño, A.; Papotti, P.; Shaar, S.; and Martino, G. D. S. 2021a. Automated Fact-Checking for Assisting Human Fact-Checkers. *arXiv/2103.07769* .

Nakov, P.; Da San Martino, G.; Elsayed, T.; Barrón-Cedeño, A.; Míguez, R.; Shaar, S.; Alam, F.; Haouari, F.; Hasanain, M.; Babulkov, N.; Nikolov, A.; Shahi, G. K.; Struß, J. M.; and Mandl, T. 2021b. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *ECIR*, 639–649.

Nakov, P.; Nayak, V.; Dent, K.; Bhatawdekar, A.; Sarwar, S. M.; Hardalov, M.; Dinkov, Y.; Zlatkova, D.; Bouchard, G.; and Augenstein, I. 2021c. Detecting Abusive Language on Online Platforms: A Critical Analysis. *arXiv/2103.00153* .

Nakov, P.; Sencar, H. T.; An, J.; and Kwak, H. 2021d. A Survey on Predicting the Factuality and the Bias of News Media. *arXiv/2103.12506* .

Pulido, C. M.; Villarejo-Carballido, B.; Redondo-Sama, G.; and Gómez, A. 2020. COVID-19 Infodemic: More Retweets for Science-Based Information on Coronavirus than for False Information. *International Sociology* 35(4): 377–392.

Qazi, U.; Imran, M.; and Ofli, F. 2020. GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information. *SIGSPATIAL Special* 12(1): 6–15.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *EMNLP*, 2931–2937.

Shahi, G. K.; and Nandini, D. 2020. FakeCovid – A Multilingual Cross-Domain Fact Check News Dataset for COVID-19. In *ICWSM Workshop*.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19(1): 22–36.

Shuja, J.; Alanazi, E.; Alasmary, W.; and Alashaikh, A. 2020. Covid-19 Open Source Data Sets: A Comprehensive Survey. *Applied Intelligence* 1–30.

Song, X.; Petrak, J.; Jiang, Y.; Singh, I.; Maynard, D.; and Bontcheva, K. 2021. Classification Aware Neural Topic Model for COVID-19 Disinformation Categorisation. *PLOS ONE* 16(2).

Tchechmedjiev, A.; Fafalios, P.; Boland, K.; Gasquet, M.; Zloch, M.; Zapilko, B.; Dietze, S.; and Todorov, K. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In *ISWC*, 309–324.

Thorne, J.; and Vlachos, A. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *COLING*, 3346–3359.

Thorne, J.; Vlachos, A.; Cocarascu, O.; Christodoulopoulos, C.; and Mittal, A. 2019. The FEVER 2.0 Shared Task. In *FEVER*, 1–6.

Vasileva, S.; Atanasova, P.; Màrquez, L.; Barrón-Cedeño, A.; and Nakov, P. 2019. It Takes Nine to Smell a Rat: Neural Multi-Task Learning for Check-Worthiness Prediction. In *RANLP*, 1229–1239.

Vidgen, B.; Hale, S.; Guest, E.; Margetts, H.; Broniatowski, D.; Waseem, Z.; Botelho, A.; Hall, M.; and Tromble, R. 2020. Detecting East Asian Prejudice on Social Media. In *Workshop on Online Abuse and Harms*, 162–172.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The Spread of True and False News Online. *Science* 359(6380): 1146–1151.

Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *ACL*, 422–426.

Zhou, X.; Mulay, A.; Ferrara, E.; and Zafarani, R. 2020. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. In *CIKM*, 3205–3212.