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

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov*, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, Preslav Nakov
Qatar Computing Research Institute, HBKU, Qatar
*Sofia University "St Kliment Ohridski", Sofia, Bulgaria
{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. Taking them into account we define an annotation schema and detailed annotation instructions, which reflect these perspectives. We performed initial annotations using this schema, and our initial experiments demonstrated sizable improvements over the baselines. Now, we issue a *call to arms* to the research community and beyond to join the fight by supporting our crowdsourcing annotation efforts.

1 Introduction

The year 2020 has 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 a long time on social media, where they find and post valuable information, a big part of which is about COVID-19. Unfortunately, amidst this rapid influx of information, there is also a spread of disinformation and harmful content in general, fighting which is of utmost importance.

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 WHO and the UN referring to it as the first global *infodemic*.

A number of initiatives were launched to fight this infodemic, primarily in social media, with focus on building large collections of tweets and then analyzing their content, source, propagators, and spread (Leng et al., 2020; Medford et al., 2020; Miller, 2020; Mourad et al., 2020; Shahi et al., 2020; Vidgen et al., 2020; Yang et al., 2020).

Most of 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.

For example, 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 1a, it should be flagged with a warning or even removed from the social media platform to prevent 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 schema further covers several categories of good posts including such containing advice (see Figure 1b), discussing action taken (see Figure 1c), calling for action, discussing possible cure, or asking a question. Such posts could be useful for journalists, policymakers, and society as a whole.

We organize the annotations with 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 a specific kind of harm to society, and (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. Keeping this in mind and in order to reduce the annotation effort and to increase the quality of the annotations, we developed a volunteer-based crowd annotation setups based on the Micromappers platform.[1]

The rest of this paper is organized as follows: Section 2 contains our call for arms. Section 3 offers a brief overview of previous work. Section 4 describes the process of data collection, the annotation instructions, and the annotation platform we use. Section 5 discusses our initial experiments and the evaluation results. Finally, Section 6 concludes and points to possible directions for future work.

2 Call to Arms

We invite everyone to join our crowdsourcing annotation efforts and to label some new tweets, thus supporting the fight against the COVID-19 infodemic. We will make all such annotations public at https://github.com/firojalam/COVID-19-tweets-for-check-worthiness.

As of present, we focus on English and Arabic tweets, but we plan extensions for other languages in the future. Here is the annotation link for English:

http://micromappers.qcri.org/project/covid19-tweet-labelling/

And here is the annotation link for Arabic:

http://micromappers.qcri.org/project/covid19-arabic-tweet-labelling/

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

(a) Bad cure

(b) Advice

(c) Action taken

Figure 1: Example of COVID-19 tweets.

3 Related Work

There have been a number of COVID-19 Twitter datasets: many without labels, other using distant supervision, and very few manually annotated.

Chen et al. (2020) built a multi-lingual dataset of 123M tweets. Abdul-Mageed et al. (2020) collected an even larger dataset, which covers more than COVID-19. Banda et al. (2020) collected 152M curated tweets. Qazi et al. (2020) built the GeoCoV19 dataset, consisting of 524M multilingual tweets, including 491M with GPS coordinates.

There are also two Arabic datasets, again without manual annotations (Alqurashi et al., 2020; Haouari et al., 2020).

Medford et al. (2020) collected tweets matching hashtags related to COVID-19 and then measured the frequency of keywords related to infection prevention practices, vaccination, and racial prejudice.

Cinelli et al. (2020) studied rumor amplification in five social media platforms, including Twitter. The rumors were labeled using distant supervision: a rumor was defined as a post that spreads an article from a questionable news source (using source labels from Media Bias Fact Check). In contrast, we have careful manual annotation and many labels.

Zhou et al. (2020) created the ReCOVery dataset, which combines news articles about COVID-19 with tweets about these articles. The articles in turn are labeled as credible vs. non-credible using distant supervision by projecting the label from their publishers, based on Media Bias/Fact Check.

Vidgen et al. (2020) studied COVID-19 prejudices against East Asians. They manually labeled a dataset of 20K tweets into four categories: hostile, criticism, prejudice, and neutral.

The closest work to ours is that of Song et al. (2020), 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. These 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).

Finally, Ding et al. (2020) have an interesting position paper discussing the challenges in combating the COVID-19 infodemic in terms of data, tools, and ethics. Other relevant work includes research on disinformation propagation (Huang and Carley, 2020; Mourad et al., 2020; Pastor-Escuredo and Tarazona, 2020; Shahi et al., 2020), studying cultural, social and political entanglements (Leng et al., 2020), and identifying disinformation campaigns (Vargas et al., 2020).

See also a recent survey: (Shuja et al., 2020).

4 Annotation Setup

In this section, we first discuss the data for the pilot annotation. Then, we present the annotation schema that we developed after a lot of analysis and discussion, and which we refined during the pilot annotations.

4.1 Data for the Pilot Annotation

We collected tweets about COVID-19 in March 2020, in English and Arabic. We then selected the most retweeted tweets for the annotation. Here are the keywords we used:

- **English:** #covid19, #CoronavirusOutbreak, #Coronavirus, #Corona, #CoronaAlert, #CoronaOutbreak, Corona, covid-19
- **Arabic:** #كورونا، كورونا (Corona), #فيروس_كورونا_الجديد, #فيروس_كورونا_المستجد (novel Coronavirus), #فيروس_كورنا، #فيروس_كورونا (Coronavirus), and #كورونا_الجديد (new Corona)

4.2 Annotation Schema and Instructions

We designed the annotation instructions after careful analysis and discussion, followed by iterative refinement in the process of pilot annotation. Our annotation schema is organized into seven questions about the input tweet. Below, we give a general idea about each question; the full annotation instructions can be found in the links in Section 2.

4.2.1 Q1: Does the tweet contain a verifiable factual claim?

This is an objective question, and it proved very easy to annotate. Positive examples include[2] tweets that state a definition, mention a quantity in the present or 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, and we ask her to answer the question, without checking anything else. This is a *Yes*/*No* question, but we also have a *Don't know or can't judge* answer, which is to be used in tricky cases, e.g., when the tweet is not in English or Arabic. If the annotator selects *Yes*, then questions 2–5 are to be answered as well; otherwise, they are skipped automatically.

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

4.2.2 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 as follows:

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*

4.2.3 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*

4.2.4 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*

4.2.5 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. This question has five answers like the previous three questions, but the answers are not on an ordinal scale; instead, they focus on the reason why there is or is not a need to fact-check the target tweet.

A. *NO, no need to check*: there is no need to fact-check the tweet, e.g., because it is not interesting, is 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 WHO, etc. An example of such a claim is as follows: "*China has 24 times more people than Italy...*"

C. *YES, not urgent*: the tweet should be fact-checked by a professional 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.

4.2.6 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

4.2.7 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. Similarly to Q5 and Q6, the answers 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 and it was labeled with some of the *YES* subcategories 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

More detailed annotation instructions with examples are provided in the annotation platform, where there is also a tutorial; see Section 2 for the links.

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, Q4) use an ordinal scale, and 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.

4.3 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,[3] which has resulted again in showing questions Q6 and Q7 only.

Using the 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 crowd-sourcing 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.

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

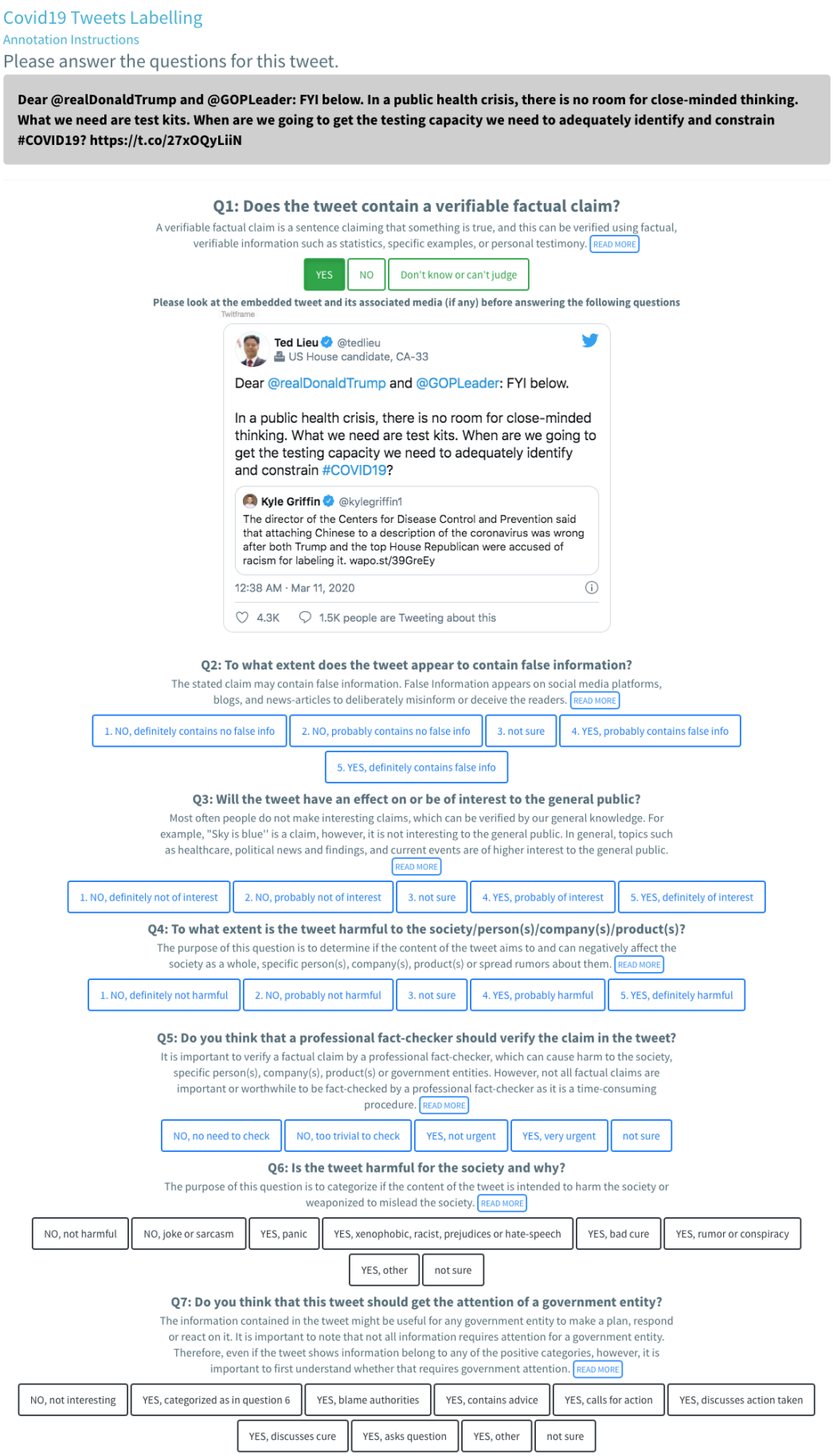

Figure 2: **The platform for an English tweet:** a *Yes* answer for Q1 has shown questions Q2–Q7 and their answers.

Figure 3: **The platform for an Arabic tweet:** a *No* answer for Q1 has only shown Q6 and Q7 only. (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...*)

4.4 Pilot Annotation Dataset

With an initial set of tweets collected, annotation guidelines developed, and annotation platforms for English and for Arabic in place, we performed pilot annotations in order to test the platform and to refine the annotation guidelines.

We annotated a total of 504 English and 218 Arabic tweets, focusing on the most retweeted tweets in our initial collection (see Section 4.1). Thus, in the English dataset, we have 504 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*. In the Arabic dataset, we have 218 tweets for Q1, Q6, and Q7, but only 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 disagreements. 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.

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 0.41–0.60, 0.61–0.80, and 0.81–1.0 correspond to moderate, substantial and perfect agreement, respectively (Landis and Koch, 1977).

5 Experiments and Evaluation

We performed some experiments on the pilot annotation dataset in order to assess to what extent it was feasible to learn from it.

We first performed standard pre-processing of the tweets: removing hash tags and other symbols, and replacing URLs and usernames by special tags.

We then explored three classifiers with diverse input representations: (*i*) SVM, which is word-based, (*ii*) FastText (Joulin et al., 2017), which uses context-independent word embeddings, and (*iii*) BERT (Devlin et al., 2019), which produces and uses contextualized word embeddings.

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.

For the SVM model, we used TF.IDF-weighted word n-grams, $n \in \{1, 2, 3\}$. We went beyond unigrams to model the context. As this yielded a large number of features, we only kept the 3,000 most frequent n-grams. We used a linear kernel.

For the FastText model, we used embeddings both for words and for character n-grams.

For the BERT-based models, we used the implementation in Hugging Face (Wolf et al., 2019). We fine-tuned `bert-base-uncased` for English and `bert-base-multilingual-uncased` for Arabic for three epochs as is common practice. Instability was an issue, and thus we performed ten reruns using different random seeds, and we selected the best model based on $\texttt{train}_{dev}$.

The evaluation results in Table 1 show that most models outperformed the majority class baseline by a sizable margin. The best model for English was BERT, which is not surprising. However, for Arabic, FastText was better; this can be attributed to its use of character n-grams, which are useful given the morphological complexity of Arabic.

	English				Arabic			
Question	Maj.	SVM	FT	BERT	Maj.	SVM	FT	mBERT
Q1	45.6	**64.8**	**72.8**	**87.6**	50.2	**72.9**	**74.4**	**88.1**
Q2	42.6	41.1	**44.0**	**48.5**	27.2	**43.3**	**47.4**	**42.8**
Q3	43.8	41.7	**48.3**	**57.6**	38.2	**49.1**	**83.1**	27.0
Q4	19.4	**41.5**	**35.5**	**41.6**	31.8	**56.4**	**54.4**	**43.7**
Q5	21.3	**37.6**	**37.6**	**50.4**	22.2	**57.4**	**77.2**	**59.0**
Q6	52.6	**50.4**	**53.9**	**57.2**	61.5	**68.6**	**79.3**	40.9
Q7	49.1	**58.6**	**57.8**	**54.6**	64.0	**69.1**	**75.7**	**66.3**
Average	39.2	**48.0**	**50.0**	**56.8**	42.1	**59.5**	**70.2**	**52.5**

Table 1: **Results for English and Arabic (weighted F1).** *Maj.* is the majority class baseline, and *FT* stands for FastText. The results that improve over the majority class baseline are shown in **bold**, and the best result for each question and language is underlined.

6 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. This is because the problem is much broader than what is typically thought of a matter of factuality: in the context of the COVID-19 infodemic, malicious content includes not only fake news, rumors, and conspiracy theories, but also the promotion of fake cures, panic, racism, xenophobia, and mistrust in the authorities, among others.

Annotating so many aspects is challenging and time-consuming. Moreover, some aspects are intrinsically subjective, which means we really need multiple annotators per example, as we have found in our preliminary manual annotations. 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 setups 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 crowdsourcing annotation efforts.

In the near future, we plan to support the annotation platforms with fresh tweets. We further plan to release annotation platforms for other languages. Last but not least, we plan regular releases of the data obtained thanks to the crowdsourcing efforts.

Acknowledgments

This research is part of the Tanbih project,[4] which aims to limit the impact of disinformation, "fake news," propaganda and media bias by making users aware of what they are reading.

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

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2020. Mega-COV: A billion-scale dataset of 65 languages for COVID-19. *arXiv:2005.06012*.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large Arabic Twitter dataset on COVID-19. *arXiv/2004.04315*.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration. *arXiv:2004.03688*.

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 Surveill*, 6(2):e19273.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *arXiv:2003.05004*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '19, pages 4171–4186, Minneapolis, MN, USA.

Kaize Ding, Kai Shu, Yichuan Li, Amrita Bhattacharjee, and Huan Liu. 2020. Challenges in combating COVID-19 infodemic – data, tools, and ethics. *arXiv:2005.13691*.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. ArCOV-19: The first Arabic COVID-19 twitter dataset with propagation networks. *arXiv:2004.05861*.

Binxuan Huang and Kathleen M. Carley. 2020. Disinformation and misinformation on Twitter during the novel coronavirus outbreak. *arXiv:2006.04278*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 427–431, Valencia, Spain.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *arXiv:1809.08193*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yan Leng, Yujia Zhai, Shaojing Sun, Yifei Wu, Jordan Selzer, Sharon Strover, Julia Fensel, Alex Pentland, and Ying Ding. 2020. Analysis of misinformation during the COVID-19 outbreak in China: cultural, social and political entanglements. *arXiv:2005.10414*.

Richard J. Medford, Sameh N. Saleh, Andrew Sumarsono, Trish M. Perl, and Christoph U. Lehmann. 2020. An "infodemic": Leveraging high-volume Twitter data to understand public sentiment for the COVID-19 outbreak. *medRxiv:10.1101/2020.04.03.20052936*.

Carl Miller. 2020. Coronavirus: Far-right spreads COVID-19 'infodemic' on Facebook. *https://www.bbc.com/news/technology-52490430*.

Azzam Mourad, Ali Srour, Haidar Harmanani, Cathia Jenainatiy, and Mohamad Arafeh. 2020. Critical impact of social networks infodemic on defeating coronavirus COVID-19 pandemic: Twitter-based study and research directions. *arXiv:2005.08820*.

David Pastor-Escuredo and Carlota Tarazona. 2020. Characterizing information leaders in Twitter during COVID-19 crisis. *arXiv:2005.07266*.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special*, 12(1):615.

Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2020. An exploratory study of COVID-19 misinformation on Twitter. *arXiv:2005.05710*.

Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. 2020. COVID-19 datasets: A survey and future challenges. *medRxiv/10.1101/2020.05.19.20107532*.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2020. Classification aware neural topic model and its application on a new COVID-19 disinformation corpus. *arXiv:2006.03354*.

Luis Vargas, Patrick Emami, and Patrick Traynor. 2020. On the detection of disinformation campaign activity with network analysis. *arXiv:2005.13466*.

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. Detecting East Asian prejudice on social media. *arXiv:2005.03909*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv:abs/1910.03771*.

Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. 2020. Prevalence of low-credibility information on Twitter during the COVID-19 outbreak. *arXiv:2004.14484*.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A multimodal repository for COVID-19 news credibility research. *arXiv:2006.05557*.