

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

Introduction

We define a comprehensive annotation schema for tweets, covering the COVID-19 pandemic, that goes beyond factuality and potential to do harm.

Covid19 Tweets Labelling

Annotation Instructions

Please answer the questions for this tweet.

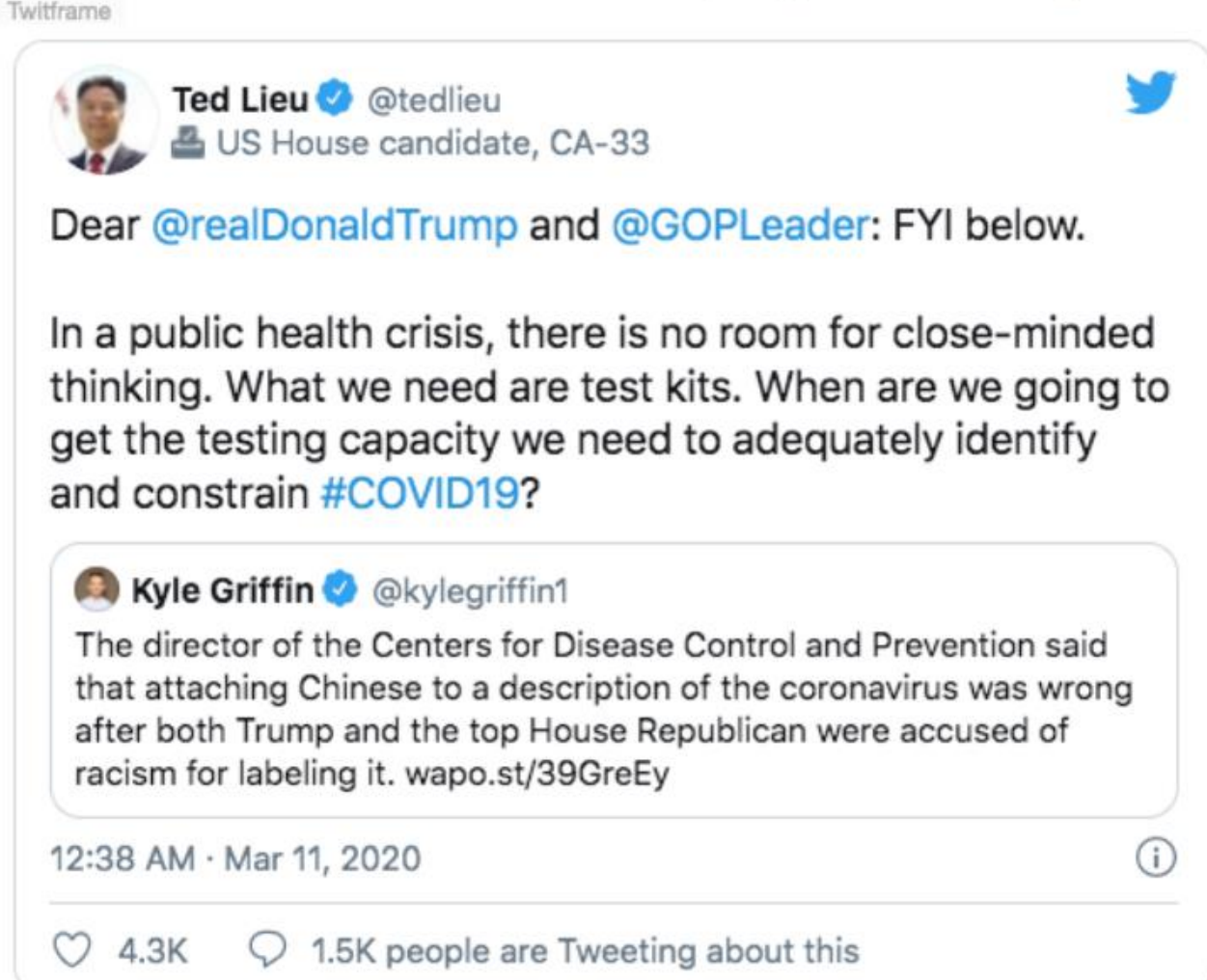
Dear @realDonaldTrump and @GOPLeader: FYI below. In a public health crisis, there is no room for close-minded thinking. What we need are test kits. When are we going to get the testing capacity we need to adequately identify and constrain #COVID19? https://t.co/27xOQyLiIn

Q1: Does the tweet contain a verifiable factual claim?

A verifiable factual claim is a sentence claiming that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. [READ MORE](#)

YES NO Don't know or can't judge

Please look at the embedded tweet and its associated media (if any) before answering the following questions



Q2: To what extent does the tweet appear to contain false information?

The stated claim may contain false information. False information appears on social media platforms, blogs, and news-articles to deliberately misinform or deceive the readers. [READ MORE](#)

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

Q3: Will the tweet have an effect on or be of interest to the general public?

Most often people do not make interesting claims, which can be verified by our general knowledge. For example, "Sky is blue" is a claim, however, it is not interesting to the general public. In general, topics such as healthcare, political news and findings, and current events are of higher interest to the general public. [READ MORE](#)

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)/product(s)?

The purpose of this question is to determine if the content of the tweet aims to and can negatively affect the society as a whole, specific person(s), company(s), product(s) or spread rumors about them. [READ MORE](#)

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

Q5: Do you think that a professional fact-checker should verify the claim in the tweet?

It is important to verify a factual claim by a professional fact-checker, which can cause harm to the society, specific person(s), company(s), product(s) or government entities. However, not all factual claims are important or worthwhile to be fact-checked by a professional fact-checker as it is a time-consuming procedure. [READ MORE](#)

NO, no need to check NO, too trivial to check YES, not urgent YES, very urgent not sure

Q6: Is the tweet harmful for the society and why?

The purpose of this question is to categorize if the content of the tweet is intended to harm the society or weaponized to mislead the society. [READ MORE](#)

NO, not harmful NO, joke or sarcasm YES, panic YES, xenophobic, racist, prejudices or hate-speech YES, bad cure YES, rumor or conspiracy YES, other not sure

Q7: Do you think that this tweet should get the attention of a government entity?

The information contained in the tweet might be useful for any government entity to make a plan, respond or react on it. It is important to note that not all information requires attention for a government entity. Therefore, even if the tweet shows information belong to any of the positive categories, however, it is important to first understand whether that requires government attention. [READ MORE](#)

NO, not interesting YES, categorized as in question 6 YES, blame authorities YES, contains advice YES, calls for action YES, discusses action taken YES, discusses cure YES, asks question YES, other not sure

Figure 1. An annotation illustration – answering Yes for Q1 reveals Q2-Q7 with their respective answers

Acknowledgments

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

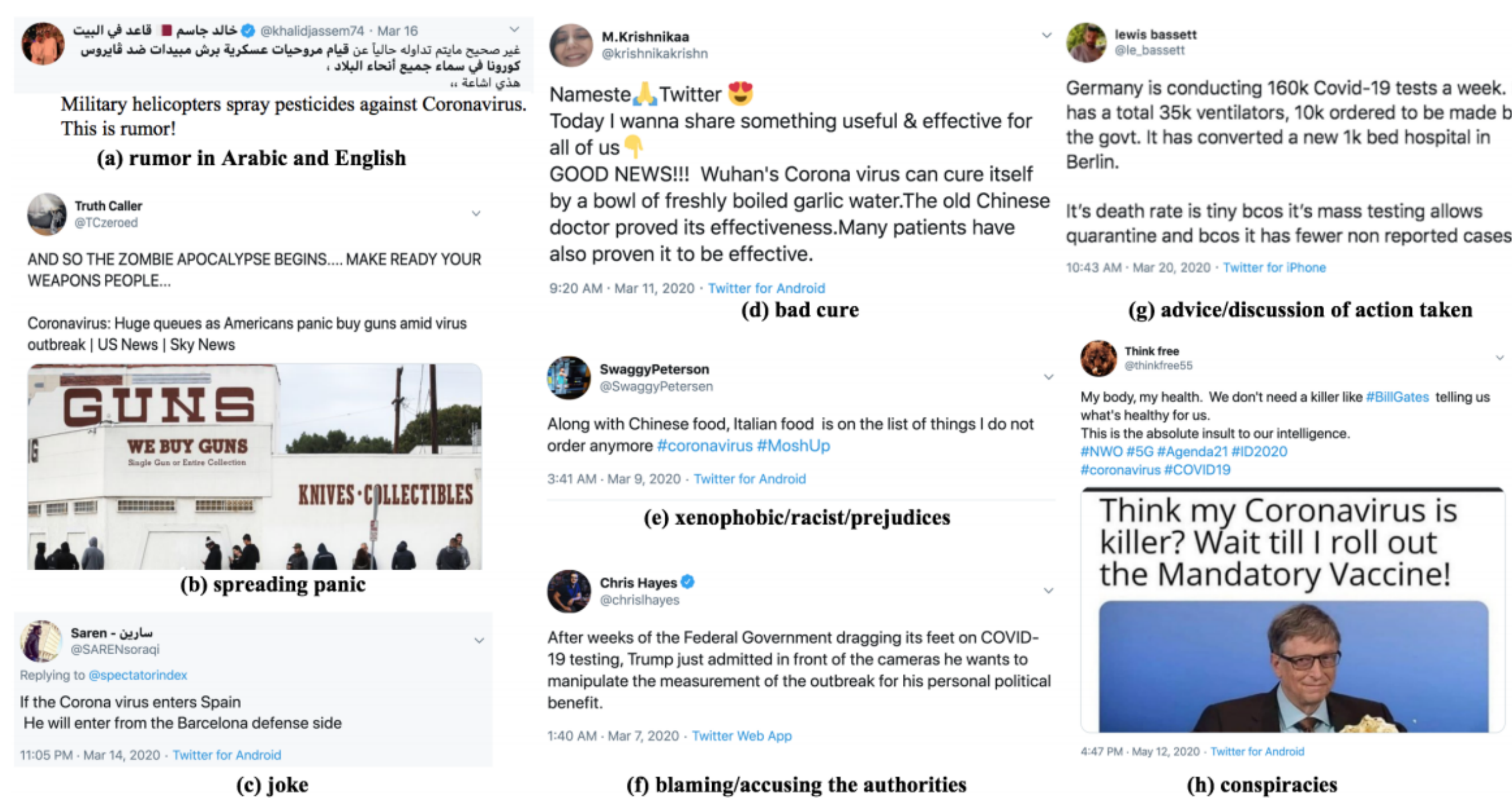


Figure 2. Example tweets, which would be of potential interest to journalists, fact-checkers, social media platforms, policy makers, government entities, and the society as a whole.

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

<https://micromappers.qcri.org/project/covid19-tweet-labelling/>

And here is the annotation link for Arabic:

<https://micromappers.qcri.org/project/covid19-arabic-tweet-labelling/>

Experiments and Evaluation

- SVM with TF-IDF weighted word n -grams, $n \in \{1,2,3\}$
- FastText with word and character based n -gram embeddings
- BERT

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

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.