

# Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society

Firoj Alam,<sup>1</sup> Shaden Shaar,<sup>1</sup> Fahim Dalvi,<sup>1</sup> Hassan Sajjad,<sup>1</sup> Alex Nikolov,<sup>2</sup>  
Hamdy Mubarak,<sup>1</sup> Giovanni Da San Martino,<sup>3</sup> Ahmed Abdelali,<sup>1</sup> Nadir Durrani,<sup>1</sup>  
Kareem Darwish,<sup>1</sup> Abdulaziz Al-Homaid,<sup>1</sup> Wajdi Zaghrouani,<sup>4</sup> Tommaso Caselli,<sup>5</sup>  
Gijs Danoe,<sup>5</sup> Friso Stolk,<sup>5</sup> Britt Bruntink<sup>5</sup> and Preslav Nakov<sup>1</sup>

<sup>1</sup> Qatar Computing Research Institute, HBKU, Qatar, <sup>2</sup> Sofia University, Sofia, Bulgaria

<sup>3</sup> University of Padova, Italy, <sup>4</sup> Hamad Bin Khalifa University, Qatar

<sup>5</sup> University of Groningen, The Netherlands

{fialam, faimaduddin, hsajjad, hmubarak, aabelali, ndurrani}@hbku.edu.qa,

{abalthomaid, kdarwish, pnakov}@hbku.edu.qa,

alexnickolow@gmail.com, dasan@math.unipd.it, wzaghrouani@hbku.edu.qa

t.caselli@rug.nl, {g.danoe, b.m.bruntink, f.r.p.stolk}@student.rug.nl

## Abstract

With the emergence of the COVID-19 pandemic, the political and the medical aspects of disinformation merged as the problem got elevated to a whole new level to become *the first global infodemic*. Fighting this infodemic has been declared one of the most important focus areas of the World Health Organization, with dangers ranging from promoting fake cures, rumors, and conspiracy theories to spreading xenophobia and panic. Addressing the issue requires solving a number of challenging problems such as identifying messages containing claims, determining their check-worthiness and factuality, and their potential to do harm as well as the nature of that harm, to mention just a few. To address this gap, we release a large dataset of 16K manually annotated tweets for fine-grained disinformation analysis that (i) focuses on COVID-19, (ii) combines the perspectives and the interests of journalists, fact-checkers, social media platforms, policy makers, and society, and (iii) covers Arabic, Bulgarian, Dutch, and English. Finally, we show strong evaluation results using pretrained Transformers, thus confirming the practical utility of the dataset in monolingual vs. multilingual, and single task vs. multitask settings.

## 1 Introduction

The rise of social media has made them one of the main channels for information dissemination and consumption. As a result, nowadays, many people rely on social media as their primary source of news (Perrin, 2015), attracted by the broader choice of information sources and by the ease for anybody to become a news producer.

Unfortunately, the democratic nature of social media has raised questions about the quality and the factuality of the information that is shared on these platforms. Eventually, social media have become one of the main channels to spread disinformation.

Figure 1 demonstrates how online users discuss topics related to COVID-19 in social media. We can see that the problem goes beyond factuality: there are tweets spreading rumors (Figure 1a), instilling panic (Figure 1b), making jokes (Figure 1c), promoting fake cures (Figure 1d), spreading xenophobia, racism, and prejudices (Figure 1e), or promoting conspiracy theories (Figure 1h).

Other examples in Figure 1 contain information that could be potentially useful and might deserve the attention of government entities. For example, the tweet in Figure 1f blames the authorities for their inaction regarding COVID-19 testing. The tweet in Figure 1g is useful both for policy makers and for the general public as it discusses action taken and suggest actions that probably should be taken elsewhere to fight the pandemic.

For the tweets in Figure 1, it is necessary to understand whether the information is correct, harmful, calling for action to be taken by relevant authorities, etc. Rapidly sorting these questions is crucial to help organizations channel their efforts, and to counter the spread of disinformation, which may cause panic, mistrust, and other problems.

Addressing these issues requires significant effort in terms of (i) defining comprehensive annotation guidelines, (ii) collecting tweets about COVID-19 and sampling from them, (iii) annotating the tweets, and (iv) training and evaluating models. Given the interconnected nature of these issues, it is more efficient to address them simultaneously.

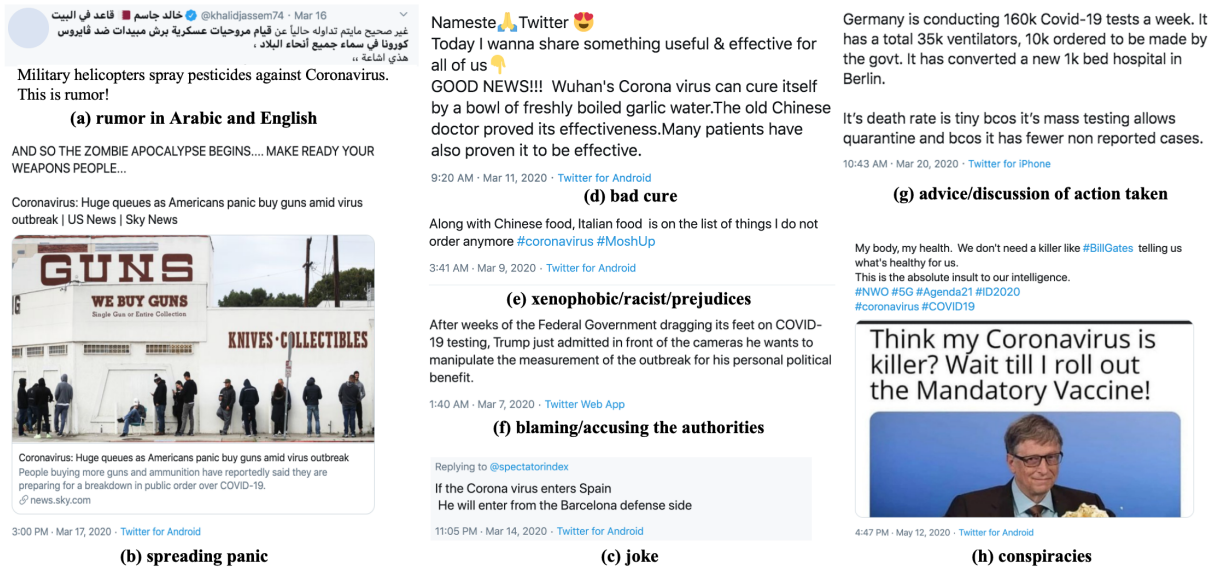


Figure 1: Examples of tweets that would be of potential interest to journalists, fact-checkers, social media platforms, policy makers, government entities, and the society as a whole.

With this consideration in mind, we adopt a *multifaceted* approach, which is motivated by engaging with different stakeholders such as journalists and policy makers. We focused on three key aspects, which are formulated into seven questions: (i) Check worthiness and veracity of the tweet (Q1-4 and Q5). (ii) Harmfulness to society (Q6); and (iii) Call for action addressing a government / policy makers (Q7). Q1–Q5 were motivated by conversations with journalists and professional fact-checkers, while Q6–Q7 were formulated in conversations with a Ministry of Public Health.

Our contributions can be summarized as follows:

- We develop a large manually annotated dataset of 16K tweets related to the COVID-19 infodemic in four languages (Arabic, Bulgarian, Dutch, and English), using a schema that combines the perspective of journalists, fact-checkers, social media platforms, policy-makers, and the society.
- We demonstrate sizable performance gains over popular deep contextualized text representations (such as BERT), when using multitask learning, cross-language learning, and when modeling the social context of the tweet, as well as the propagandistic nature of the language used.
- We make our data and code freely available.<sup>1</sup>

<sup>1</sup><https://github.com/firojalam/COVID-19-disinformation>

## 2 Related Work

**Fact-Checking** Research on fact-checking claims is largely based on datasets mined from major fact-checking organizations. Some of the larger datasets include the *Liar, Liar* dataset of 12.8K claims from PolitiFact (Wang, 2017), the *ClaimsKG* dataset and system (Tchechmedjiev et al., 2019) of 28K claims from eight 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* dataset (Rashkin et al., 2017). There have been also datasets for other languages, created in a similar fashion, e.g., for Arabic (Baly et al., 2018; Alhindi et al., 2021).

A number of datasets were created as part of shared tasks. In most cases, they performed their own annotation, either (a) manually, e.g., the SemEval tasks on determining the veracity of rumors (Derczynski et al., 2017; Gorrell et al., 2019), propaganda detection in news articles and memes (Da San Martino et al., 2020a; Dimitrov et al., 2021a,b), fact-checking in community question answering forums (Mihaylova et al., 2019), the CLEF CheckThat! lab on identification and verification of claims (Nakov et al., 2018; Elsayed et al., 2019; Barrón-Cedeño et al., 2020; Shaar et al., 2020; Nakov et al., 2021c; Shaar et al., 2021b,c), or (b) using crowdsourcing, e.g., the FEVER task on fact extraction and verification, focusing on claims about Wikipedia content (Thorne et al., 2018, 2019).

Unlike our work, the above datasets did not focus on tweets (they used claims from news, speeches, political debates, community question answering fora, or were just made up by human annotators; RumourEval is a notable exception), targeted factuality only (we cover a number of other issues), were limited to a single language (typically English; except for CLEF), and did not focus on COVID-19.

**Check-Worthiness Estimation** Another relevant research line is on detecting check-worthy claims in political debates using manual annotations (Hassan et al., 2015) or by observing the selection of fact-checkers (Gencheva et al., 2017; Patwari et al., 2017; Jaradat et al., 2018; Vasileva et al., 2019).

**COVID-19 Research** There are a number of COVID-19 Twitter datasets: some unlabeled (Chen et al., 2020; Banda et al., 2021; Haouari et al., 2021), some automatically labeled with location information (Abdul-Mageed et al., 2021; Qazi et al., 2020), some labeled using distant supervision (Cinelli et al., 2020; Zhou et al., 2020), and some manually annotated (Song et al., 2020; Vidgen et al., 2020; Shahi and Nandini, 2020; Pulido et al., 2020; Dharawat et al., 2020).

There is also work on credibility (Cinelli et al., 2020; Pulido et al., 2020; Zhou et al., 2020), racial prejudices and fear (Medford et al., 2020; Vidgen et al., 2020), as well as situational information, e.g., caution and advice (Li et al., 2020), as well as on detecting mentions and stance with respect to known misconceptions (Hossain et al., 2020).

The closest work to ours is that of Song et al. (2020), who collected false and misleading claims about COVID-19 from IFCN Poynter, and annotated them as (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. These categories partially overlap with ours, but account for less perspectives. Moreover, we cover both true and false claims, we focus on tweets (while they have general claims), and we cover four languages.

Last but not least, we have described the general *annotation schema* in previous work (Alam et al., 2021a). Unlike that work, here we focus on the *dataset*, which is much larger and covers four languages, and we present a rich set of experiments.

## 3 Dataset

### 3.1 Data Collection

We collected tweets by specifying a target language (English, Arabic, Bulgarian, or Dutch), a set of COVID-19 related keywords, as shown in Figure 2, and different time frames: from January 2020 till March 2021. We collected original tweets (no retweets or replies), we removed duplicates using a similarity-based approach (Alam et al., 2021b), and we filtered out tweets with less than five words. Finally, we selected the most frequently liked and retweeted tweets for annotation.

**English:** #covid19, #CoronavirusOutbreak, #Coronavirus, #Corona, #CoronaAlert, #CoronaOutbreak, Corona, covid-19, COVID vaccine, Covid-19 vaccine, #covidvaccine, corona vaccine, #vaccinate, #vaccine, vaccine  
**Arabic:** كورونا, كورونا (Corona), فيروس كورونا الجديد, #فيروس\_كورونا\_المستجد, (novel Coronavirus), #فيروس\_كورونا, and #Coronavirus, لقاح, مطعوم, تطعيم, كورونا\_الجديد, لقاءات,  
**Bulgarian:** #корона, #коронавирус, коронавирус, корона  
**Dutch:** #coronavirus, #COVID19, #coronaviruschina, #coronavirusNederland, #Italië, #RIVM, #coronavirusnederland, #CoronavirusOutbreak, #COVID-19, #CoronaVirusUpdates, #persconferentie, #Vindicat, #COVID\_19, #hamsteren, #coronagekte, #coronapocalypse, #coronadebat, #scholendicht, #COVID19NL, #samentegencorona, #StayHomeSaveLives, #thuisonderwijs, #thuisblijven, #thuiswerken, #ikleesthuis, #groepimmunitet, #LockdownNow, #blijfthuis, #houdafstand, #anderhalvemeter, #testen

Figure 2: The keywords used to collect the tweets.

### 3.2 Annotation Task

The annotation task consists of determining whether a tweet contains a factual claim, as well as its veracity, its potential to cause harm (to the society, to a person, to an organization, or to a product), whether it needs verification, and how interesting it is for policy makers. These are then formulated into seven questions presented in Table 1.

The full annotation instructions we gave to the annotators, together with examples, can be found in Appendix D. To facilitate the annotation task, we used the annotation platform described in Alam et al. (2021a). There were 10, 14, 5, and 4 annotators for English, Arabic, Bulgarian, and Dutch, respectively. We used three annotators per tweet, native speakers or fluent in the respective language, male and female, with qualifications ranging from undergrads to PhDs in various disciplines. We resolved the cases of disagreement in a consolidation discussion including external consolidators.

Table 3 shows two tweets, annotated for all questions. The first tweet contains a harmful factual claim with a causal argument of interest to the public and requiring urgent fact-checking. Moreover, it appears to spread rumors. It also attacks government officials, and thus might need the attention of government entities. The second tweet contains a non-harmful factual claim of interest to the general public, which is probably true, but should be fact-checked urgently. It might be of interest to policy makers as it discusses protection from COVID-19.

### 3.3 Labels

The annotation was designed in a way that the fine-grained multiclass labels can be easily transformed into binary labels by mapping all *Yes\** into **Yes**, and all *No\** into **No**, and dropping the *not sure* tweets.

Although some of the questions are correlated (for Q1-Q5, this is on purpose), the annotation instructions are designed, so that the dataset can be used independently for different tasks. Questions Q2-Q4 (see Table 1) can be seen as categorical or numerical (i.e., on a Likert scale), and thus can be addressed in a classification or in an ordinal regression setup. Below, we will use classification.

### 3.4 Statistics

We annotated a total of 4,542, 4,966, 3,697, and 2,665 tweets for English, Arabic, Bulgarian, and Dutch, respectively. Table 1 shows the distribution of the class labels for all languages.

The distribution for Q1 is quite balanced: 64% *Yes* vs. 36% *No*. Only tweets that contain factual claims were annotated for Q2–Q5.

For question Q2, 81% of the tweets were judged to contain no false information, for 6% the judges were unsure, and 13% were suspected to possibly contain false information. Note that this is not fact-checking, but just a subjective judgment about whether the claim seems credible.

Exp.	Class labels	En	Ar	Bg	Nl
<b>Q1: Does the tweet contain a verifiable factual claim?</b>		<b>4,542</b>	<b>4,966</b>	<b>3,697</b>	<b>2,665</b>
Bin	No	1,651	1,527	1,130	1,412
	Yes	2,891	3,439	2,567	1,253
<b>Q2: To what extent does the tweet appear to contain false information?</b>		<b>2,891</b>	<b>3,439</b>	<b>2,567</b>	<b>1,253</b>
Multi	No, definitely contains no false info	222	137	102	190
	No, probably contains no false info	2,272	2,465	2,166	718
	not sure	213	22	219	113
	Yes, probably contains false info	142	764	5	162
	Yes, definitely contains false info	42	51	75	70
Bin	No	2,494	2,602	2,268	908
	Yes	184	815	80	232
<b>Q3: Will the tweet’s claim have an impact on or be of interest to the general public?</b>		<b>2,891</b>	<b>3,439</b>	<b>2,567</b>	<b>1,253</b>
Multi	No, definitely not of interest	11	9	2	108
	No, probably not of interest	94	120	68	181
	not sure	8	14	0	21
	Yes, probably of interest	2,481	2,047	2,000	645
	Yes, definitely of interest	297	1249	497	298
Bin	No	105	129	70	289
	Yes	2,778	3,296	2,497	943
<b>Q4: To what extent does the tweet appear to be harmful to the society, a person(s), a company(s) or a product(s)?</b>		<b>2,891</b>	<b>3,439</b>	<b>2,567</b>	<b>1,253</b>
Multi	No, definitely not harmful	1,107	1,591	437	520
	No, probably not harmful	1,126	1,088	1,876	449
	not sure	21	22	17	23
	Yes, probably harmful	505	433	196	204
	Yes, definitely harmful	132	305	41	57
Bin	No	2,233	2,233	2,313	969
	Yes	637	637	237	261
<b>Q5: Do you think that a professional fact-checker should verify the claim in the tweet?</b>		<b>2,891</b>	<b>3,439</b>	<b>2,567</b>	<b>1,247</b>
Multi	No, no need to check	472	163	721	410
	No, too trivial to check	1,799	1,948	1,326	330
	Yes, not urgent	513	1086	422	309
	Yes, very urgent	107	242	98	198
Bin	No	2,271	2,111	2,047	740
	Yes	620	1,328	520	507

Table 1: Statistics about Q1–Q5. In rows with a question, the number refers to the total number of tweets for the respective language. Bin: binary, Multi: multiclass.

For Q3, which asks whether the tweet is of potential interest to the general public, the distribution is quite skewed towards *Yes*: 94% of the examples. This can be attributed to the fact that we selected the tweets based on frequency of retweets and likes, and these would be the interesting tweets.

For Q4, which asks whether the tweet is harmful to the society, we can see that the labels vary widely from not harmful to harmful; yet, most are not harmful.

Exp.	Class labels	En	Ar	Bg	Nl
<b>Q6: Is the tweet harmful to the society and why?</b>					
	No, joke or sarcasm	95	155	200	162
	No, not harmful	4,040	3,872	3,017	2,254
	not sure	2	12	4	9
Multi	Yes, bad cure	4	6	7	10
	Yes, other	33	23	4	7
	Yes, panic	90	347	305	35
	Yes, rumor conspiracy	246	425	151	159
	Yes, xenophobic racist prejudices or hate speech	32	126	9	29
Bin	No	4,135	4,027	3,217	2,416
	Yes	405	927	476	240
<b>Q7: Do you think that this tweet should get the attention of a government entity?</b>					
	No, not interesting	3,892	1,598	3,186	2,092
	not sure	6	12	0	4
	Yes, asks question	7	129	1	116
Multi	Yes, blame authorities	181	93	51	177
	Yes, calls for action	63	61	8	43
	Yes, classified as in question 6	249	725	333	136
	Yes, contains advice	18	102	10	50
	Yes, discusses action taken	35	695	25	32
	Yes, discusses cure	60	1,536	79	8
	Yes, other	31	15	4	7
Bin	No	3,892	1,598	3,186	2,092
	Yes	644	3,356	511	569

Table 2: Statistics about Q6–Q7.

For Q5, which asks whether a professional fact-checker should verify the claim, the majority of the cases were either *Yes, not urgent* (23%) or *No, no need to check* (17%). It appears that a professional fact-checker should verify the claim urgently in a relatively small number of cases (6%).

For questions Q2–4, the *not sure* cases are very rare. However, they are substantially more prevalent for Q2 (6%), which is hard to annotate, as in many cases, it requires access to external information. When annotating Q2 (as well as Q3–Q7, but not Q1), the annotators were presented the tweet as it appears in Twitter, which allows them to see some context, e.g., the user identifier, a snapshot of linked webpage, a video, an image, etc.

For Q6, most of the tweets were considered *not harmful* for the society or a *joke*. However, 1% of the tweets were found to be *xenophobic, racist, prejudices or hate speech*, 6% to be *rumor conspiracy*, and 5% to be *spreading panic*.

For Q7, the vast majority of the tweets were not interesting for policy makers and government entities. However, 3% blamed the authorities.

**Tweet 1:** *This is unbelievable. It reportedly took Macron’s threat to close the UK border for Boris Johnson to finally shutdown bars and restaurants. The Elysee refers to UK policy as ‘benign neglect’. This failure of leadership is costing lives.*

Q1: Yes  
Q2: No, probably contains no false info  
Q3: Yes, probably of interest  
Q4: Yes, definitely harmful  
Q5: Yes, very urgent  
Q6: Yes, rumor, or conspiracy  
Q7: Yes, blames authorities

**Tweet 2:** *An antiviral spray against novel #coronavirus has developed in Shanghai Public Health Clinical Center, which can be put into throat as shield from virus. The spray can greatly help protect front-line medical staff, yet mass-production for public use is not available for now. <https://t.co/bmRzCssCY5>*

Q1: Yes  
Q2: not sure  
Q3: Yes, definitely of interest  
Q4: No, definitely not harmful  
Q5: Yes, very urgent  
Q6: No, not harmful  
Q7: Yes, discusses cure

Table 3: Examples of annotated English tweets.

### 3.5 Inter-Annotation Agreement

We assessed the quality of the annotations by computing inter-annotator agreement. As mentioned earlier, three annotators independently annotated each tweet, following the provided annotation instructions, and the cases of disagreement were resolved in a consolidation discussion including external consolidators. We computed the Fleiss Kappa ( $\kappa$ ) between each annotator and the consolidated label, using (a) the original multiclass labels, and (b) binary labels. The results for the English dataset are shown in Table 4, where we can see that overall, there is moderate to substantial agreement.<sup>2</sup> The Kappa value is higher for objective questions such as Q1, and it is lower for subjective and partially subjective questions;<sup>3</sup> the number of labels is also a factor. The agreement for the other languages is also moderate to substantial for all questions and also both for binary and for multiclass labels; see Appendix E for more detail.

<sup>2</sup>Recall 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).

<sup>3</sup>Our agreement is much higher than for related tasks (Roitero et al., 2020): Krippendorff’s  $\alpha$  in [0.066; 0.131].

Agree. Pair	Q1	Q2	Q3	Q4	Q5	Q6	Q7
<b>Multiclass</b>							
A1 - C	0.81	0.73	0.59	0.74	0.79	0.67	0.73
A2 - C	0.67	0.53	0.44	0.45	0.39	0.65	0.47
A3 - C	0.78	0.58	0.63	0.61	0.70	0.17	0.42
<b>Avg</b>	<b>0.75</b>	<b>0.61</b>	<b>0.55</b>	<b>0.60</b>	<b>0.63</b>	<b>0.50</b>	<b>0.54</b>
<b>Binary</b>							
A1 - C	0.81	0.73	0.77	0.85	0.84	0.77	0.92
A2 - C	0.67	0.58	0.53	0.43	0.52	0.33	0.57
A3 - C	0.78	0.70	0.63	0.70	0.74	0.11	0.57
<b>Avg</b>	<b>0.75</b>	<b>0.67</b>	<b>0.64</b>	<b>0.66</b>	<b>0.70</b>	<b>0.40</b>	<b>0.69</b>

Table 4: Inter-annotator agreement using Fleiss Kappa ( $\kappa$ ) for the English dataset. *A* refers to annotator, and *C* refers to consolidation.

## 4 Experimental Setup

We experimented with binary and multiclass settings for all languages, using deep contextualized text representations based on large-scale pretrained transformer models such as BERT, mBERT, RoBERTa, XLM-R, etc. We further performed multitask and cross-language learning, and we modeled the social context of the tweet, as well as the propagandistic nature of the language used.

### 4.1 Data Preprocessing

The preprocessing includes removal of hash-symbols and non-alphanumeric symbols, case folding, URL replacement with a URL tag, and username replacement with a user tag. We generated a stratified split (Sechidis et al., 2011) of the data into 70%/10%/20% for training/development/testing. We used the development set to tune the model hyper-parameters.

**Models** Large-scale pretrained Transformer models have achieved state-of-the-art performance for several NLP tasks. We experimented with several such models to evaluate their efficacy under various training scenarios such as, binary vs. multiclass classification, multilingual setup, etc.

We used BERT (Devlin et al., 2019) and RoBERTa for English, AraBERT (Antoun et al., 2020) for Arabic, and BERTje (de Vries et al., 2019) for Dutch. We further used multilingual transformers such as (Liu et al., 2019), multilingual BERT (mBERT) and XLM-r (Conneau et al., 2020). Finally, we used static embeddings from FastText (Joulin et al., 2017).

For Transformer models, we used the Transformer toolkit (Wolf et al., 2020). We fine-tuned each model using the default settings for ten epochs as described in (Devlin et al., 2019). Due to instability, we performed ten reruns for each experiment using different random seeds, and we picked the model that performed best on the development set.

For FastText, we used embeddings pretrained on Common Crawl, which were released by FastText for different languages.

### 4.2 Multitask Learning

While question Q1, Q2, . . . , Q7 can be deemed as independent tasks, some questions are interrelated and information in one can help improve the predictive performance for another task. For example, Q5 asks whether the claim in a tweet should be checked by a professional fact-checker. A tweet is more likely to be worth fact-checking if its factuality is under question (Q2), if it is interesting for the general public (Q3), and, more importantly, if it is harmful (Q4). This interdependence between the tasks (which was by design) motivated multitask learning with the goal of improving the performance of the classifier on Q5 using Q2, Q3, and Q4 as auxiliary tasks. We applied multitask learning by aggregating task-specific dense layers of transformers. More specifically, for the four questions, we computed the cross-entropy loss for each task independently and we then combined them linearly:  $L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4$  where the lambdas sum up to 1.

### 4.3 Twitter/Propagandistic/Botometer Features

Previous work has demonstrated the utility of modeling the social context for related tasks such as predicting factuality (Canini et al., 2011; Baly et al., 2020), and thus we extracted context features from the Twitter object. We further modeled the degree of propagandistic content in the tweet, and we also used bot-related features.

The features from the Twitter object include general information about the tweet’s content, as well as about its author, i.e., whether the account is verified, whether it uses the default profile picture, the number of years since the account’s creation, the number of followers, statuses, and friends, whether the tweet contains quotes, media or a URL, and the factuality of the website it points to.<sup>4</sup>

<sup>4</sup>From <http://mediabiasfactcheck.com>

		English				Arabic				Bulgarian				Dutch			
Q.	Cls.	Maj.	FT	BT	RT	Maj.	FT	ArBT	XLM-r	Maj.	FT	mBT	XLM-r	Maj.	FT	BTje	XLM-r
<b>Binary (Coarse-grained)</b>																	
Q1	2	48.7	<b>77.7</b>	<b>76.5</b>	<b>78.6</b>	56.8	<b>63.1</b>	<b>83.8</b>	<b>84.2</b>	58.3	<b>75.5</b>	<b>84.0</b>	<b>87.6</b>	36.5	<b>61.9</b>	<b>75.4</b>	<b>80.0</b>
Q2	2	91.6	89.0	<b>92.1</b>	<b>92.7</b>	68.3	<b>81.7</b>	<b>84.0</b>	<b>83.1</b>	95.0	85.2	94.7	<b>95.0</b>	64.9	<b>87.9</b>	<b>75.1</b>	<b>83.1</b>
Q3	2	96.3	69.3	<b>96.4</b>	<b>96.9</b>	96.3	82.0	96.0	<b>96.3</b>	96.5	79.3	96.0	<b>96.5</b>	62.3	<b>69.9</b>	<b>76.9</b>	<b>78.3</b>
Q4	2	66.7	<b>96.3</b>	<b>85.6</b>	<b>89.0</b>	67.2	<b>96.2</b>	<b>90.3</b>	<b>89.0</b>	86.8	<b>96.5</b>	<b>87.7</b>	<b>88.4</b>	63.9	<b>72.7</b>	<b>77.1</b>	<b>83.9</b>
Q5	2	67.7	<b>83.8</b>	<b>80.6</b>	<b>84.4</b>	46.8	<b>74.0</b>	<b>65.9</b>	<b>66.7</b>	70.5	<b>81.5</b>	<b>80.5</b>	<b>82.9</b>	44.4	<b>75.3</b>	<b>66.8</b>	<b>70.9</b>
Q6	2	86.7	<b>92.1</b>	<b>88.9</b>	<b>90.5</b>	72.5	<b>79.3</b>	<b>88.9</b>	<b>89.8</b>	83.2	<b>95.0</b>	<b>84.5</b>	<b>85.1</b>	84.7	74.9	<b>86.9</b>	<b>88.1</b>
Q7	2	78.3	<b>80.6</b>	<b>85.5</b>	<b>86.1</b>	57.7	<b>81.6</b>	<b>77.4</b>	<b>77.4</b>	80.1	<b>87.2</b>	<b>81.6</b>	<b>81.7</b>	65.6	<b>74.1</b>	<b>78.3</b>	<b>79.6</b>
Avg.		76.6	<b>84.1</b>	<b>86.5</b>	<b>88.3</b>	66.5	<b>79.7</b>	<b>83.8</b>	<b>83.7</b>	81.5	<b>85.8</b>	<b>87.0</b>	<b>88.2</b>	60.3	<b>73.8</b>	<b>76.6</b>	<b>80.5</b>
<b>Multiclass (Fine-grained)</b>																	
Q2	5	67.9	44.7	<b>69.2</b>	<b>70.6</b>	62.9	53.3	<b>75.6</b>	<b>76.2</b>	77.3	<b>78.8</b>	<b>77.8</b>	<b>79.3</b>	36.5	<b>39.7</b>	<b>45.7</b>	<b>51.1</b>
Q3	5	78.9	57.4	<b>82.5</b>	<b>82.8</b>	44.4	<b>75.6</b>	<b>53.7</b>	<b>59.5</b>	64.2	<b>78.2</b>	<b>68.1</b>	<b>68.8</b>	32.0	<b>77.7</b>	<b>50.9</b>	<b>53.9</b>
Q4	5	19.9	<b>69.2</b>	<b>56.0</b>	<b>58.0</b>	28.1	<b>54.2</b>	<b>46.9</b>	<b>50.6</b>	58.8	<b>69.0</b>	<b>65.6</b>	<b>67.1</b>	21.0	<b>42.9</b>	<b>46.3</b>	<b>53.1</b>
Q5	5	46.8	<b>84.9</b>	<b>62.0</b>	<b>70.0</b>	41.2	<b>52.6</b>	<b>52.6</b>	<b>52.4</b>	36.0	<b>81.5</b>	<b>58.0</b>	<b>61.6</b>	18.4	<b>69.6</b>	<b>40.7</b>	<b>46.4</b>
Q6	8	84.0	71.7	<b>86.5</b>	<b>87.7</b>	68.7	<b>71.5</b>	<b>82.2</b>	<b>84.8</b>	76.6	<b>79.6</b>	<b>77.2</b>	<b>78.8</b>	74.4	46.0	<b>76.7</b>	<b>76.3</b>
Q7	10	78.1	<b>82.4</b>	<b>83.4</b>	<b>85.3</b>	13.8	<b>40.8</b>	<b>57.5</b>	<b>61.6</b>	80.1	<b>66.8</b>	<b>81.7</b>	<b>81.8</b>	65.4	45.3	<b>72.2</b>	<b>74.1</b>
Avg.		62.6	<b>68.4</b>	<b>73.3</b>	<b>75.8</b>	43.2	<b>58.0</b>	<b>61.4</b>	<b>64.2</b>	65.5	<b>75.6</b>	<b>71.4</b>	<b>72.9</b>	41.3	<b>53.5</b>	<b>55.4</b>	<b>59.1</b>

Table 5: **Monolingual experiments.** We report weighted  $F_1$  for binary (top) and multiclass (bottom) experiments for English, Arabic, Bulgarian, and Dutch using various Transformers and FastText (FT). 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. **BT**: BERT, **ArBT**: Monolingual BERT in Arabic (AraBERT), **RT**: RoBERTa. **mBT**: multilingual BERT, **BTje**: Monolingual BERT in Dutch (BERTje), **XLM-r**: XLM-RoBERTa.

The propagandistic features include two scores modeling the degree to which the message is propagandistic: one from the Propy (Barrón-Cedeño et al., 2019; Barrón-Cedeño et al., 2019) and one from the Prta (Da San Martino et al., 2020b) systems, as implemented in Tanbih (Zhang et al., 2019).

We extracted bot-related features using the Botometer (Davis et al., 2016). This includes a score about whether the tweet author is likely to be a bot, as well as content-, network- and friend-related scores. These features are summarized in Appendix (Table 9).

#### 4.4 Baseline

For all tasks, we use a majority class baseline. Note that for questions with highly imbalanced class distribution, this baseline could be very high, which can make it hard for models to improve upon (see Table 5). For example, in the Arabic dataset for Q3 in the binary setting, the tweets from the *Yes* category comprise 96% of the total.

#### 4.5 Evaluation Measures

We report weighted  $F_1$  score, which takes into account class imbalance. In Appendix C, we further report some other evaluation measures such as accuracy and macro-average  $F_1$  score.

## 5 Evaluation Results

### 5.1 Binary Classification

The evaluation results for binary classification are shown in the first half of Table 5.

**English** Most models outperformed the baseline. RoBERTa outperformed the other models in five of the seven tasks, and FastText was best on the remaining two.

**Arabic** In all the cases except for Q3 (which has a very skewed distribution as we mentioned above), all models performed better than the baseline. The strongest models were FastText and XLM-r, each winning 3 of the seven tasks. AraBERT was best on one of the tasks.

**Bulgarian** For Bulgarian, most models outperformed the baselines. We also have a highly imbalanced distribution for Q2 (96.6% ‘No’) and for Q3 (97.3% ‘Yes’), which made for a very hard to beat baseline. XLM-r was best for four out of seven tasks, and FastText was best on the remaining three.

**Dutch** For Dutch, all models managed to outperform the majority class baseline, except for FastText on Q6 (due to class imbalance). XLM-r performed best in five out of the seven tasks, and FastText was best on the other two.

## 5.2 Multiclass Classification

The bottom part of Table 5 shows the multiclass results. The *Cls* column shows the number of classes per task. We can see that this number ranges in [5,10], and thus the multiclass setup is a much harder compared to binary classification. This explains the much lower results compared to the binary case (including for the baseline).

**English** Most models outperformed the baseline. The most successful model was RoBERTa, which was best for four out of the six tasks; FastText was best on the remaining two tasks.

**Arabic** Almost all models outperformed the majority class baseline for all tasks (except for FastText on Q2). FastText was best for three of the six tasks, XLM-r was best on two, and AraBERT was best on the remaining one.

**Bulgarian** All models outperformed the baselines for all tasks. FastText was best for four tasks, and XLM-r was best for the remaining two.

**Dutch** Most models outperformed the majority class baseline. XLM-r was best for three of the six tasks, FastText was best on two, and BERTje won the remaining one.

## 5.3 Discussion

Overall, the experimental results above have shown that there is no single model that performs universally best across all languages, all tasks, and all class sizes. We should note, however, the strong performance of RoBERTa for English, and of XLM-r for the remaining languages.

Interestingly, language-specific models, such as AraBERT for Arabic and BERTje for Dutch, were not as strong as multilingual ones such as XLM-r. This could be partially explained by the fact that for them we used a base-sized models, while for XLM-r we used a large model.

Finally, we should note the strong performance of context-free models such as FastText. We believe that it is suitable for the noisy text of tweets due to its ability to model not only words but also character  $n$ -grams. In future work, we plan to try transformers specifically trained on tweets and/or on COVID-19 related data such as BERTweet (Nguyen et al., 2020) and COVID-Twitter-BERT (Müller et al., 2020).

## 6 Advanced Experiments

Next, we performed some additional, more advanced experiments, including multilingual training, modeling the Twitter context, the use of propagandistic language, and whether the user is likely to be a bot, as well as multitask learning. We describe each of these experiments in more detail below.

### 6.1 Multilingual Training

We experimented with a multilingual setup, where we combined the data from all languages. We fine-tuned a multilingual model (mBERT),<sup>5</sup> separately for each question. The results are shown in Table 6, where the *Mul* columns shows the multilingual fine-tuning results, which are to be compared to the monolingual fine-tuning results in the previous respective columns. We can see that the differences are small and that the results are mixed. Multilingual fine-tuning helps a bit in about half of the cases, but it also hurts a bit in the other half of the cases. This is true both in the binary and in the multiclass setting.

		English	Arabic	Bulgarian	Dutch		
Q.	Cls.	EN	AR	BG	NL	Mul	Mul
<b>Binary (Coarse-grained)</b>							
Q1	2	76.5	<b>77.5</b>	<b>82.6</b>	81.5	<b>84.0</b>	81.8
Q2	2	92.1	<b>92.6</b>	<b>81.4</b>	78.8	<b>94.7</b>	94.4
Q3	2	<b>96.4</b>	<b>96.4</b>	96.1	<b>96.5</b>	96.0	<b>96.5</b>
Q4	2	<b>85.6</b>	83.9	<b>87.7</b>	87.2	<b>87.7</b>	87.2
Q5	2	<b>80.6</b>	78.6	63.1	<b>66.5</b>	80.5	<b>83.2</b>
Q6	2	<b>88.9</b>	85.6	84.6	<b>85.6</b>	84.5	<b>85.6</b>
Q7	2	<b>85.5</b>	79.9	73.4	<b>79.9</b>	<b>81.6</b>	79.9
<b>Avg.</b>		86.5	84.9	81	82.4	87.5	87.8
<b>Multiclass (Fine-grained)</b>							
Q2	5	69.2	<b>70.2</b>	70.8	<b>72.0</b>	<b>77.8</b>	<b>77.8</b>
Q3	5	82.5	<b>82.9</b>	55.8	<b>55.9</b>	68.1	<b>68.3</b>
Q4	5	56.0	<b>56.3</b>	<b>48.2</b>	43.8	65.6	<b>68.9</b>
Q5	5	<b>62.0</b>	61.2	<b>56.0</b>	54.6	<b>58.0</b>	56.3
Q6	8	<b>86.5</b>	84.8	<b>79.0</b>	78.9	77.2	<b>77.8</b>
Q7	10	<b>83.4</b>	<b>83.4</b>	<b>54.7</b>	53.5	<b>81.7</b>	80.2
<b>Avg.</b>		73.3	73.1	60.7	59.8	71.4	71.5
						55.3	54.9

Table 6: **Multilingual experiments using mBERT.** Shown are results for monolingual vs. multilingual models (weighted  $F_1$ ). *Mul* is trained on the combined English, Arabic, Bulgarian, and Dutch data.

<sup>5</sup>We also tried XLM-r, but it performed worse.



## 6.2 Twitter/Propagandistic/Botometer

We conducted experiments with Twitter, propaganda, and botness features alongside the posteriors from the BERT classifier, which we combined using XGBoost (Chen and Guestrin, 2016). The results are shown in Table 7. We can see that many of the combinations yielded improvements, with botness being the most useful, followed by propaganda, and finally by the Twitter object features.

Binary (Coarse-grained)						
Q.	Cls	BERT	B+TF	B+Prop	B+Bot	B+All
Q1	2	76.5	<b>76.9</b>	<b>77.1</b>	<b>77.8</b>	<b>76.8</b>
Q2	2	92.1	91.8	<b>92.3</b>	<b>92.3</b>	<b>92.4</b>
Q3	2	<u>96.4</u>	96.3	<u>96.4</u>	<u>96.4</u>	96.3
Q4	2	85.6	<b>86.5</b>	<b>86.5</b>	<b>86.7</b>	<b>86.4</b>
Q5	2	80.6	<b>82.0</b>	<b>81.5</b>	<b>81.9</b>	<b>81.4</b>
Q6	2	88.9	88.9	<b>89.6</b>	<b>89.4</b>	87.6
Q7	2	85.5	84.1	<b>85.6</b>	<b>86.2</b>	83.9
Multiclass (Fine-grained)						
Q2	5	69.2	<b>69.4</b>	<b>70.0</b>	<b>70.3</b>	69.1
Q3	5	82.5	81.2	82.2	82.2	81.6
Q4	5	56.0	52.7	55.9	<b>56.8</b>	53.4
Q5	4	62.0	60.9	<b>63.2</b>	<b>62.8</b>	58.2
Q6	8	86.5	84.3	86.4	<b>86.6</b>	84.1
Q7	10	83.4	79.6	<b>83.7</b>	<b>83.9</b>	80.8

Table 7: **Experiments with social features and BERT (weighted  $F_1$ )**. Improvements over BERT (**B**) are shown in **bold**, while the highest scores for each question are underlined. **TF**: Tweet features, **Prop**: propaganda features, **Bot**: Botometer features.

## 6.3 Multitask Learning

For the multitask learning experiments, we used BERT and RoBERTa on the English dataset, in a multiclass setting, fine-tuned with a multiclass objective on Q2–Q5. The results are shown in Table 8. We achieved sizable improvements for Q2, Q4, and Q5 over the single-task setup. However, performance degraded for Q3, probably due to the skewed label distribution for this question.

English, multiclass				
	BERT(S)	BERT(M)	RoBERTa(S)	RoBERTa(M)
Q2	69.2	<b>72.9</b>	70.62	<b>73.85</b>
Q3	<b>82.5</b>	71.6	<b>82.84</b>	67.34
Q4	56.0	<b>67.9</b>	58.04	<b>66.95</b>
Q5	62.0	<b>76.8</b>	70.02	<b>75.75</b>

Table 8: **Multitask learning experiments** (weighted  $F_1$ ). **S**: Single task, **M**: Multitask.

## 7 Conclusion and Future Work

We presented a large manually annotated dataset of COVID-19 tweets, aiming to help in the fight against the COVID-19 infodemic. The dataset combines the perspectives and the interests of journalists, fact-checkers, social media platforms, policy-makers, and society as a whole. It includes tweets in Arabic, Bulgarian, Dutch, and English, and we are making it freely available to the research community. We further reported a number of evaluation results for all languages using various transformer architectures. Moreover, we performed advanced experiments, including multilingual training, modeling the Twitter context, the use of propagandistic language, and whether the user is likely to be a bot, as well as multitask learning.

In future work, we plan to explore multimodality and explainability (Yu et al., 2021). We further want to model the task as a multitask ordinal regression (Baly et al., 2019), as Q2–Q5 are defined on an ordinal scale. Moreover, we would like to put the data and the system in some practical use; in fact, we have already used them to analyze disinformation about COVID-19 in Bulgaria (Nakov et al., 2021a) and Qatar (Nakov et al., 2021b). Finally, the data will be used in a shared task at the CLEF-2022 CheckThat! lab; part of it was used for the NLP4IF-2021 shared task (Shaar et al., 2021a).

## Acknowledgments

We thank Akter Fatema, Al-Awthman Ahmed, Al-Dobashi Hussein, El Messelmani Jana, Fayoumi Sereen, Mohamed Esraa, Ragab Saleh, and Shurafa Chereen for helping with the Arabic annotations.

We also want to thank the Atlantic Club in Bulgaria and DataBee for their support for the Bulgarian annotations.

This research is part of the Tanbih mega-project, 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.

This material is also based upon work supported by the US National Science Foundation under Grants No. 1704113 and No. 1828199.

This publication was also partially made possible by the innovation grant No. 21 – Misinformation and Social Networks Analysis in Qatar from Hamad Bin Khalifa University’s (HBKU) Innovation Center. The findings achieved herein are solely the responsibility of the authors.

## Ethics Statement

### Dataset Collection

We collected the dataset using the Twitter API<sup>6</sup> with keywords that only use terms related to COVID-19, without other biases. We followed the terms of use outlined by Twitter.<sup>7</sup> Specifically, we only downloaded public tweets, and we only distribute dehydrated Twitter IDs.

### Biases

We note that some of the annotations are subjective, and we have clearly indicated in the text which these are. Thus, it is inevitable that there would be biases in our dataset. Yet, we have a very clear annotation schema and instructions, which should reduce biases.

### Misuse Potential

Most datasets compiled from social media present some risk of misuse. We, therefore, ask researchers to be aware that our dataset can be maliciously used to unfairly moderate text (e.g., a tweet) that may not be malicious based on biases that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required in order to ensure this does not occur.

### Intended Use

Our dataset can enable automatic systems for analysis of social media content, which could be of interest to practitioners, professional fact-checker, journalists, social media platforms, and policymakers. Such systems can be used to alleviate the burden for social media moderators, but human supervision would be required for more intricate cases and in order to ensure that the system does not cause harm.

Our models can help fight the infodemic, and they could support analysis and decision making for the public good. However, the models can also be misused by malicious actors. Therefore, we ask the potential users to be aware of potential misuse. With the possible ramifications of a highly subjective dataset, we distribute it for research purposes only, without a license for commercial use. Any biases found in the dataset are unintentional, and we do not intend to do harm to any group or individual.

<sup>6</sup><http://developer.twitter.com/en/docs>

<sup>7</sup><http://developer.twitter.com/en/developer-terms/agreement-and-policy>

## References

- Muhammad Abdul-Mageed, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2021. [Mega-COV: A billion-scale dataset of 100+ languages for COVID-19](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '21, pages 3402–3420, Online. Association for Computational Linguistics.
- Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2021a. [Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '21, pages 913–922.
- Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021b. [CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '21, pages 923–932.
- Gerald Albaum. 1997. The likert scale revisited. *Market Research Society. Journal.*, 39(2):1–21.
- Tariq Alhindi, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. [AraStance: A multi-country and multi-domain dataset of Arabic stance detection for fact checking](#). In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF '21, pages 57–65, Online. Association for Computational Linguistics.
- Gordon W Allport and Leo Postman. 1947. *The psychology of rumor*. Henry Holt.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. [AraBERT: Transformer-based model for Arabic language understanding](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, OSACT '20, pages 9–15, Marseille, France. European Language Resource Association.
- Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. [MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.
- Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. [What was written vs. who](#)

- read it: [News media profiling using text analysis and social media context](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3364–3374, Online. Association for Computational Linguistics.
- Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. [Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 2109–2116, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. [Integrating stance detection and fact checking in a unified corpus](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 21–27, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. [A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration](#). *Epidemiologia*, 2(3):315–324.
- Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. [Proppy: A system to unmask propaganda in online news](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, pages 9847–9848, Honolulu, Hawaii, USA. AAAI.
- Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. [Check-that! at CLEF 2020: Enabling the automatic identification and verification of claims in social media](#). In *Proceedings of the 42nd European Conference on Information Retrieval*, ECIR '19, pages 499–507, Lisbon, Portugal. Springer.
- Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. [Proppy: Organizing the news based on their propagandistic content](#). *Information Processing and Management*, 56(5):1849–1864.
- David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. 2018. [Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate](#). *American journal of public health*, 108(10):1378–1384.
- Kevin R Canini, Bongwon Suh, and Peter L Pirolli. 2011. [Finding credible information sources in social networks based on content and social structure](#). In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, SocialCom/PASSAT '11*, pages 1–8, Boston, Massachusetts, USA.
- 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.
- Tianqi Chen and Carlos Guestrin. 2016. [XGBoost: A scalable tree boosting system](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, San Francisco, California, USA. Association for Computing Machinery.
- Matteo Cinelli, Walter Quattrociochi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. [The COVID-19 social media infodemic](#). *Scientific Reports*, 10(1):1–10.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 8440–8451, Online. Association for Computational Linguistics.
- Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. [SemEval-2020 task 11: Detection of propaganda techniques in news articles](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1377–1414. International Committee for Computational Linguistics.
- Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020b. [Prta: A system to support the analysis of propaganda techniques in the news](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online. Association for Computational Linguistics.
- Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. [BotOrNot: A system to evaluate social bots](#). In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16, pages 273–274, Montréal, Québec, Canada. International World Wide Web Conferences.
- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. [BERTje: A Dutch BERT model](#). *ArXiv:1912.09582*.

- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.
- 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: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. [Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation](#). *arXiv:2010.08743*.
- Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. [Detecting propaganda techniques in memes](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617, Online. Association for Computational Linguistics.
- Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. [SemEval-2021 task 6: Detection of persuasion techniques in texts and images](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, pages 70–98, Online. Association for Computational Linguistics.
- Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. [CheckThat! at CLEF 2019: Automatic identification and verification of claims](#). In *Advances in Information Retrieval*, ECIR '19, pages 309–315, Cologne, Germany. Springer International Publishing.
- Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. [A context-aware approach for detecting worth-checking claims in political debates](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '17, pages 267–276, Varna, Bulgaria. INCOMA Ltd.
- Genevieve Gorrell, Ahmet Aker, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2019. [SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. [ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, ANLP '21, pages 82–91, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. [Detecting check-worthy factual claims in presidential debates](#). In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 1835–1838, Melbourne, Australia. Association for Computing Machinery.
- 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*, Online. Association for Computational Linguistics.
- Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. [Claim-Rank: Detecting check-worthy claims in Arabic and English](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-HLT '18, pages 26–30, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. [Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection](#). *Digital Threats: Research and Practice*, 2(2).
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Lifang Li, Qingpeng Zhang, Xiao Wang, Jun Zhang, Tao Wang, Tian-Lu Gao, Wei Duan, Kelvin Kam-fai Tsoi, and Fei-Yue Wang. 2020. [Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo](#). *IEEE Transactions on Computational Social Systems*, 7(2):556–562.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *ArXiv:1907.11692*.
- 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 Early Public Sentiment for the Coronavirus Disease 2019 Outbreak](#). *Open Forum Infectious Diseases*, 7(7). Ofaa258.
- Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. [SemEval-2019 task 8: Fact checking in community question answering forums](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval ’19*, pages 860–869, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Martin Müller, Marcel Salathé, and Per Egil Kummerfeld. 2020. [COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter](#). *arXiv:2005.07503*.
- Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021a. [COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP ’21*, pages 1001–1013, Online. INCOMA Ltd.
- Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021b. [A second pandemic? Analysis of fake news about COVID-19 vaccines in Qatar](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP ’21*, pages 1014–1025, Online. INCOMA Ltd.
- Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. [Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims](#). In *CLEF, Lecture Notes in Computer Science*, pages 372–387, Avignon, France. Springer.
- Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021c. [Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association*, LNCS (12880). Springer.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. [BERTweet: A pre-trained language model for English tweets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP ’20*, pages 9–14, Online. Association for Computational Linguistics.
- Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. [TATHYA: A multi-classifier system for detecting check-worthy statements in political debates](#). In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM ’17*, page 2259–2262, Singapore. Association for Computing Machinery.
- Andrew Perrin. 2015. Social media usage. *Pew research center*, pages 52–68.
- Cristina M Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. 2020. [COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information](#). *International Sociology*, 35(4):377–392.
- 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):6–15.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. [Truth of varying shades: Analyzing language in fake news and political fact-checking](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP ’17*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.
- Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. [Can the crowd identify misinformation objectively? The effects of judgment scale and assessor’s background](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’20*, pages 439–448, Virtual Event, China. Association for Computing Machinery.
- Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. [On the stratification of multi-label data](#). In *Machine Learning and Knowledge Discovery in Databases, ECML-PKDD ’11*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021a. [Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection](#). In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF ’21’*, pages 82–92, Online. Association for Computational Linguistics.

- Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, Giovanni Da San Martino, Tamer Elsayed, and Preslav Nakov. 2021b. [Overview of the CLEF-2021 CheckThat! lab task 2 on detecting previously fact-checked claims in tweets and political debates](#). In *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*, CLEF '21, Bucharest, Romania (online). CEUR-WS.
- Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mucahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Javier Beltrán, Tamer Elsayed, and Preslav Nakov. 2021c. [Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates](#). In *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*, CLEF '2021, Bucharest, Romania (online). CEUR-WS.org.
- Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeño, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, and Preslav Nakov. 2020. [Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media](#). In *Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CEUR-WS.org.
- Gautam Kishore Shahi and Durgesh Nandini. 2020. [FakeCovid – a multilingual cross-domain fact check news dataset for COVID-19](#). In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, Online.
- Xingyi Song, Johann Petrak, Ye Jiang, Iknor 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*.
- Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. [ClaimsKG: A knowledge graph of fact-checked claims](#). In *Proceedings of the 18th International Semantic Web Conference*, ISWC '19, pages 309–324, Auckland, New Zealand.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 809–819, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. [The FEVER2.0 shared task](#). In *Proceedings of the Second Workshop on Fact Extraction and VERification*, FEVER '19, pages 1–6, Hong Kong, China. Association for Computational Linguistics.
- Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. [It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP'19, pages 1229–1239, Varna, Bulgaria. INCOMA Ltd.
- Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. [Detecting East Asian prejudice on social media](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, ALW '20, pages 162–172, Online. Association for Computational Linguistics.
- William Yang Wang. 2017. [“Liar, liar pants on fire”: A new benchmark dataset for fake news detection](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '20, pages 38–45, Online. Association for Computational Linguistics.
- Seunghak Yu, Giovanni Da San Martino, Mitra Mohitarami, James Glass, and Preslav Nakov. 2021. [Interpretable propaganda detection in news articles](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '21, pages 1601–1609, Online.
- Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Hae-woon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhev, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. [Tanbih: Get to know what you are reading](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, EMNLP-IJCNLP '19, pages 223–228, Hong Kong, China.
- Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. [ReCOVeRY: A multimodal repository for COVID-19 news credibility research](#). In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 3205–3212. Association for Computing Machinery.

## Appendix

### A Experimental Setup

#### A.1 Transformer Parameters

Below, we list the values of the hyper-parameters that we used for fine-tuning the Transformer models we used. We further release all our scripts, together with the data.

- Batch size: 32;
- Learning rate (Adam):  $2e-5$ ;
- Number of epochs: 10;
- Max seq length: 128.

#### Models and Number of Parameters:

- **BERT** (bert-base-uncased):  $L=12$ ,  $H=768$ ,  $A=12$ , total parameters: 110M; where  $L$  is the number of layers (i.e., Transformer blocks),  $H$  is the hidden size, and  $A$  is the number of self-attention heads;
- **RoBERTa** (roberta-base): similar to BERT-base, but with a higher number of parameters (125M);
- **AraBERT** (bert-base-arabert): same number as BERT (110M);
- **BERTje** (bert-base-dutch-cased): same number as BERT (110M);
- **RoBERTa for Bulgarian** (roberta-base-bulgarian):  $L=12$ ,  $H=768$ ,  $A=12$ , parameters=125M;
- **BERT Multilingual** (bert-base-multilingual-uncased) (mBERT): similar to BERT-base with a higher number of parameters (172M);
- **XLM-RoBERTa** (xlm-roberta-base):  $L=12$ ,  $H=768$ ,  $A=12$ ; the total number of parameters is 270M.

#### A.2 FastText Parameters

We release all the FastText parameters with our released packages. We have not listed them here due to the length of the resulting list.

#### A.3 XGBoost Parameters

We used XGBoost to run experiments with Twitter, Propaganda, Botometer, and BERT model predictions. We release the scripts with our code repository, which contains detailed the parameter settings.

#### A.4 Computing Infrastructure and Runtime

We used a server with NVIDIA Tesla V100-SXM2-32 GB GPU, 56 cores, and 256GB CPU memory. To perform an experiment for a question, on average the computing time took 40 minutes using the BERT base model. This means about four hours for all seven questions using one Transformer architecture.

## B Twitter/Propagandistic/Botometer Features Types

In the additional experiments in Section 6, we extracted features from the Twitter object, botness scores from the Botometer API, and propaganda scores from the Tanbih API. We have already described the experiments with these features in Section 6, but we did not have enough space in the main text of the paper to describe the features themselves. Table 9 aims to address this. It lists the features, offers a brief description for each one, and specifies its type, which can be one of the following:

- *Boolean features* take a value of either 0 or 1; we use them directly.
- *Categorical features* take a fixed number of possible values, and we encode them using one-hot representation.
- *Numerical features* are continuous and may take an infinite number of values; we transform the value  $x$  of such a feature according to the formula  $x' = \ln(x + 1)$ .

Tweet-Specific		Description
URL	B	Is there a URL included in the tweet?
Reply	B	Is the tweet a reply?
Quotes	B	Is this a quoted tweet?
URL	B	Does the tweet contain a URL?
Media	B	Does the tweet contain media?
Source	C	Tools/devices used to post the tweet, as an HTML-formatted string.
Domain	C	Domain of the included URL.
Num media	N	Number of media mentioned in the tweet.
Media type	C	Type of included media, e.g., image. A label (unknown, high, mixed, or low) for factuality of the linked information, e.g., if it is a news medium.
Fact	C	
User-Specific		Description
Statuses	N	Number of tweets (incl. retweets) posted.
Followers	N	The number of followers.
Friends	N	The number of following.
Favorites	N	The number of liked tweets.
Listed	N	The number of subscriptions to public lists.
Default profile	B	Has the user altered the theme or the background of the profile?
Profile img	B	Has the user uploaded a profile image?
Verified	B	Is it a verified account?
Protected	B	Has the user chosen to protect their tweets?
GEO-enabled	B	Is geotagging enabled?
Botometer		Description
Content	N	Score of the length of tweets and frequency of part-of-speech tags.
Network	N	Score about retweets, mentions, and hashtags that a user tweeted in the past.
Temporal	N	Score about time patterns of tweets.
Sentiment	N	Score about the sentiment of the user.
Friend	N	Score about users that liked or retweeted tweets by the user.
Language	C	Language used.
User	N	Score about the number of followers' user name, and consistency of shared language between the tweets.
Propaganda		Description
Prta	N	Sentence-level Prta propaganda score
Proppy	N	Article-level Proppy propaganda score

Table 9: Features modeling social context, botness, and propaganda. The middle column shows the type of feature: B is Boolean, C is categorical, and N is numerical.



## C Detailed Result by Language

The main text of the paper had only weighted F1 scores; here we report also accuracy (Acc) and macro-F1 (M-F1) for English, Arabic, Bulgarian and Dutch are shown in Tables 10, 11, 12 and 13, respectively.

Q	Binary			Multiclass		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1
<b>Majority</b>						
Q1	63.0	38.7	48.7			
Q2	94.3	48.5	91.6	77.7	17.5	67.9
Q3	97.6	49.4	96.3	85.5	18.4	78.9
Q4	76.8	43.4	66.7	36.9	10.8	19.9
Q5	77.5	43.7	67.7	61.5	19.0	46.8
Q6	91.0	47.6	86.7	89.1	11.8	84.0
Q7	85.1	46.0	78.3	85.0	9.2	78.1
<b>FastText</b>						
Q1	79.3	65.9	77.7			
Q2	90.8	61.3	89.0	46.3	28.9	44.7
Q3	69.5	66.9	69.3	65.0	33.5	57.4
Q4	97.0	54.5	96.3	78.0	20.7	69.2
Q5	85.4	65.2	83.8	89.4	14.3	84.9
Q6	93.0	58.8	92.1	72.9	68.7	71.7
Q7	81.9	71.3	80.6	86.8	23.6	82.4
<b>BERT</b>						
Q1	76.8	74.5	76.5			
Q2	92.8	60.1	92.1	73.0	25.5	69.2
Q3	97.2	54.8	96.4	85.2	27.0	82.5
Q4	85.9	79.3	85.6	56.4	40.3	56.0
Q5	81.5	71.0	80.6	64.8	37.2	62.0
Q6	90.2	62.3	88.9	88.3	22.2	86.5
Q7	87.0	68.5	85.5	85.2	27.7	83.4
<b>RoBERTa</b>						
Q1	78.8	76.8	78.6			
Q2	93.2	63.6	92.7	71.1	37.9	70.6
Q3	97.6	60.5	96.9	83.3	33.2	82.8
Q4	89.1	84.5	89.0	58.7	43.9	58.0
Q5	84.7	77.4	84.4	71.4	51.3	70.0
Q6	91.4	68.6	90.5	88.7	26.2	87.7
Q7	86.7	71.3	86.1	86.1	33.7	85.3

Table 10: Classification results on the **test set for English** using various models including a majority class baseline for different questions. Acc. is Accuracy, M-F1 is macro F1, and W-F1 is weighted average F1.

Q	Binary			Multiclass		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1
<b>Majority</b>						
Q1	69.4	41.0	56.8			
Q2	78.0	43.8	68.3	74.0	17.0	62.9
Q3	97.5	49.4	96.3	59.5	14.9	44.4
Q4	77.1	43.5	67.2	45.2	12.4	28.1
Q5	61.5	38.1	46.8	56.9	18.1	41.2
Q6	81.0	44.7	72.5	78.2	11.0	68.7
Q7	70.1	41.2	57.7	29.9	4.6	13.8
<b>FastText</b>						
Q1	64.4	60.0	63.1			
Q2	84.5	66.6	81.7	57.0	30.4	53.3
Q3	82.6	78.2	82.0	81.1	22.7	75.6
Q4	97.2	49.3	96.2	56.6	21.0	54.2
Q5	75.9	67.2	74.0	54.7	27.7	52.6
Q6	81.5	67.2	79.3	75.3	26.6	71.5
Q7	83.7	71.5	81.6	43.7	28.0	40.8
<b>AraBERT</b>						
Q1	84.1	80.7	83.8			
Q2	84.7	75.7	84.0	78.1	30.6	75.6
Q3	96.5	53.0	96.0	54.4	22.9	53.7
Q4	90.4	86.3	90.3	47.6	34.0	46.9
Q5	66.3	63.7	65.9	53.3	34.7	52.6
Q6	89.2	81.4	88.9	82.8	32.3	82.2
Q7	77.8	72.6	77.4	57.8	37.3	57.5
<b>XLm-RoBERTa</b>						
Q1	84.6	81.0	84.2			
Q2	84.0	74.4	83.1	78.7	31.4	76.2
Q3	97.5	49.4	96.3	60.6	23.7	59.5
Q4	89.1	84.3	89.0	52.1	36.5	50.6
Q5	67.1	64.5	66.7	55.3	31.7	52.4
Q6	89.8	83.3	89.8	85.1	36.4	84.8
Q7	77.8	72.6	77.4	61.7	40.8	61.6

Table 11: Classification results on the **test set for Arabic** using various models including a majority class baseline for different questions. Acc. is Accuracy, M-F1 is macro F1, and W-F1 is weighted average F1.

Q	Binary			Multiclass		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1
<b>Majority</b>						
Q1	70.5	41.4	58.3			
Q2	96.6	49.1	95.0	84.4	18.3	77.3
Q3	97.7	49.4	96.5	75.0	21.4	64.2
Q4	91.1	47.7	86.8	70.9	16.6	58.8
Q5	79.6	44.3	70.5	52.4	17.2	36.0
Q6	88.6	47.0	83.2	84.0	11.4	76.6
Q7	86.4	46.4	80.1	86.4	10.3	80.1
<b>FastText</b>						
Q1	78.8	58.1	75.5			
Q2	88.7	55.9	85.2	84.0	16.6	78.8
Q3	80.6	74.0	79.3	78.5	73.5	78.2
Q4	97.7	49.4	96.5	76.1	26.5	69.0
Q5	86.4	51.7	81.5	86.4	13.6	81.5
Q6	96.6	49.1	95.0	85.2	27.7	79.6
Q7	91.1	49.7	87.2	73.6	24.2	66.8
<b>mBERT</b>						
Q1	84.5	80.3	84.0			
Q2	96.0	49.0	94.7	80.9	27.5	77.8
Q3	96.5	49.1	96.0	71.1	27.8	68.1
Q4	87.8	62.0	87.7	68.2	26.8	65.6
Q5	81.7	68.2	80.5	59.5	41.9	58.0
Q6	86.1	58.1	84.5	80.3	16.2	77.2
Q7	82.9	58.1	81.6	84.4	17.8	81.7
<b>XLm-RoBERTa</b>						
Q1	88.0	84.7	87.6			
Q2	96.6	49.1	95.0	83.6	28.6	79.3
Q3	97.7	49.4	96.5	71.3	28.6	68.8
Q4	88.8	63.2	88.4	67.4	32.2	67.1
Q5	83.6	72.7	82.9	63.0	44.7	61.6
Q6	86.0	61.1	85.1	79.2	23.2	78.8
Q7	82.1	60.5	81.7	84.4	18.5	81.8

Table 12: Classification results on the **test set for Bulgarian** using various models including a majority class baseline for different questions. Acc. is Accuracy, M-F1 is macro F1, and W-F1 is weighted average F1.

Q	Binary			Multiclass		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1
<b>Majority</b>						
Q1	52.8	34.6	36.5			
Q2	75.4	43.0	64.9	52.8	13.8	36.5
Q3	73.5	42.4	62.3	48.8	13.1	32.0
Q4	74.7	42.8	63.9	38.1	11.0	21.0
Q5	59.5	37.3	44.4	35.3	10.4	18.4
Q6	89.6	47.3	84.7	82.4	11.3	74.4
Q7	76.0	43.2	65.6	75.8	8.6	65.4
<b>FastText</b>						
Q1	63.1	59.7	61.9			
Q2	89.8	62.6	87.9	40.1	29.7	39.7
Q3	69.9	69.8	69.9	81.8	26.6	77.7
Q4	75.9	61.9	72.7	47.2	27.1	42.9
Q5	76.2	65.1	75.3	74.5	15.6	69.6
Q6	77.6	63.1	74.9	52.0	28.2	46.0
Q7	77.6	62.2	74.1	47.2	29.4	45.3
<b>BERTje</b>						
Q1	75.5	75.3	75.4			
Q2	76.3	64.9	75.1	51.6	27.8	45.7
Q3	78.7	68.5	76.9	53.2	36.7	50.9
Q4	78.8	67.8	77.1	48.0	29.7	46.3
Q5	67.1	65.3	66.8	40.9	30.3	40.7
Q6	88.9	59.7	86.9	80.5	16.7	76.7
Q7	78.8	69.5	78.3	75.5	19.1	72.2
<b>XLm-RoBERTa</b>						
Q1	80.0	80.0	80.0			
Q2	84.2	75.9	83.1	56.7	31.2	51.1
Q3	79.1	71.1	78.3	56.3	38.4	53.9
Q4	84.1	78.4	83.9	54.4	36.1	53.1
Q5	71.0	69.7	70.9	46.8	35.0	46.4
Q6	89.1	65.5	88.1	80.7	15.7	76.3
Q7	79.4	72.3	79.6	77.0	21.3	74.1

Table 13: Classification results on the **test set for Dutch** using various models including a majority class baseline for different questions. Acc. is Accuracy, M-F1 is macro F1, and W-F1 is weighted average F1.

## D Detailed Annotation Instructions

### General Instructions:

1. For each tweet, the annotator needs to read the text, including the hashtags, and also to look at the tweet itself when necessary by going to the link (i.e., for Q2-7 it might be required to open the tweet link). The reason for not going to the tweet link for Q1 is that we wanted to reduce the complexity of the annotation task and to focus on the content of the tweet only. As for Q2, it might be important to check whether the tweet was posted by an authoritative source, and thus it might be useful for the annotator to open the tweet to get more context. After all, this is how real users perceive the tweet. Since the annotators would open the tweet's link for Q2, they can use that information for the rest of the questions as well (even though this is not required).
2. The annotators should assume the time when the tweet was posted as a reference when making judgments, e.g., *“Trump thinks, that for the vast majority of Americans, the risk is very, very low.”* would be true when he made the statement but false by the time annotations were carried out for this tweet.
3. The annotators may look at the images, the videos and the Web pages that the tweet links to, as well as at the tweets in the same thread when making a judgment, if needed.
4. The annotators are not asked to complete questions Q2-Q5 if the answer to question Q1 is **NO**.

### D.1 Verifiable Factual Claim

**Question 1:** 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. Factual claims include the following:<sup>8</sup>

- Stating a definition;
- Mentioning quantity in the present or the past;
- Making a verifiable prediction about the future;

<sup>8</sup>Inspired by (Konstantinovskiy et al., 2021).

- Statistics or specific examples;
- Personal experience or statement (e.g., *“I spent much of the last decade working to develop an #Ebola treatment.”*)
- Reference to laws, procedures, and rules of operation;
- References (e.g., URL) to images or videos (e.g., *“This is a video showing a hospital in Spain.”*);
- Statements that can be technically classified as questions, but in fact contain a verifiable claim based on the criteria above (e.g., *“Hold on - #China Communist Party now denying #CoronavirusOutbreak originated in China? This after Beijing’s catastrophic mishandling of the virus has caused a global health crisis?”*)
- Statements about correlation or causation. Such a correlation or causation needs to be explicit, i.e., sentences like *“This is why the beaches haven’t closed in Florida. <https://t.co/8x2tcQeg21>”* is not a claim because it does not explicitly say why, and thus it is not verifiable.

Tweets containing personal opinions and preferences are not factual claims. Note that if a tweet is composed of multiple sentences or clauses, at least one full sentence or clause needs to be a claim in order for the tweet to contain a factual claim. If a claim exists in a sub-sentence or a sub-clause, then the tweet is not considered to contain a factual claim. For example, *“My new favorite thing is Italian mayors and regional presidents LOSING IT at people violating quarantine”* is not a claim – it is in fact an opinion. However, if we consider *“Italian mayors and regional presidents LOSING IT at people violating quarantine”* it would be a claim. In addition, when answering this question, annotators should not open the tweet URL. Since this is a binary decision task, the answer of this question consists of two labels as defined below.

### Labels:

- **YES**: if it contains a verifiable factual claim;
- **NO**: if it does not contain a verifiable factual claim;

- **Don't know or can't judge:** the content of the tweet does not provide enough information to make a judgment. It is recommended to categorize the tweet using this label when the content of the tweet is not understandable at all. For example, it uses a language (i.e., non-English) or references difficult to understand;

### Examples:

1. *Please don't take hydroxychloroquine (Plaquenil) plus Azithromycin for #COVID19 UNLESS your doctor prescribes it. Both drugs affect the QT interval of your heart and can lead to arrhythmias and sudden death, especially if you are taking other meds or have a heart condition.*

**Label: YES**

**Explanation:** There is a claim in the text.

2. *Saw this on Facebook today and it's a must read for all those idiots clearing the shelves #coronavirus #toiletpapercrisis #auspol*

**Label: NO**

**Explanation:** There is no claim in the text.

## D.2 False Information

**Question 2:** To what extent does the tweet appear to contain false information?

The stated claim may contain false information. This question labels the tweets with the categories mentioned below. *False Information* appears on social media platforms, blogs, and news-articles to deliberately misinform or deceive readers.

**Labels:** The labels for this question are defined on a five point Likert scale (Albaum, 1997). A higher value means that it is more likely to be false:

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**

To answer this question, it is recommended to open the link of the tweet and to look for additional information to determine the veracity of the claims it makes. For example, if the tweet contains a link to an article from a reputable information source (e.g., Reuters, Associated Press, France Press, Aljazeera English, BBC), then the answer could be "... contains no false info". Note that answering this question is not required if the answer to Question 1 is **NO**.

### Examples:

1. *"Dominican Republic found the cure for Covid-19 <https://t.co/1CfA162Lq3>"*

**Label: 5. YES, definitely contains false information**

**Explanation:** This is not correct information at the time of this tweet is posted.

2. *This is Dr. Usama Riaz. He spent past weeks screening and treating patients with Corona Virus in Pakistan. He knew there was no PPE. He persisted anyways. Today he lost his own battle with coronavirus but he gave life and hope to so many more. KNOW HIS NAME*

 <https://t.co/ftSwhLCPmx>

**Label: 2. NO, probably contains no false info**

**Explanation:** The content of the tweet states correct information.

## D.3 Interest to the General Public

**Question 3:** Will the tweet's claim have an impact 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, though "*The sky is blue*" is a claim, it is not interesting to the general public. In general, topics such as healthcare, political news, and current events are of higher interest to the general public. Using the five point Likert scale the labels are defined below.

**Labels:** The labels are on a 5-point Likert scale:

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**

## Examples:

1. *Germany is conducting 160k Covid-19 tests a week. It has a total 35k ventilators, 10k ordered to be made by the govt. It has converted a new 1k bed hospital in Berlin. It's death rate is tiny bcos it's mass testing allows quarantine and bcos it has fewer non reported cases.*

**Label: 4. YES: probably of interest**

**Explanation:** This information is relevant and of high interest for the general population as it reports how a country deals with COVID-19.

2. *Fake news peddler Dhruv Rathee had said: "Corona virus won't spread outside China, we need not worry" Has this guy ever spoke something sensible? <https://t.co/siBAwIR8Pn>*

**Label: 2. NO, probably not of interest**

**Explanation:** The information is not interesting for the general public as it is an opinion and discusses the statement by someone else.

## D.4 Harmfulness

**Question 4:** To what extent does the tweet appear to be harmful to society, person(s), company(s) or product(s)? The purpose of this question is to determine whether the content of the tweet aims to and can negatively affect society as a whole, specific person(s), company(s), product(s), or spread rumors about them. The content intends to harm or *weaponize the information*<sup>9</sup> (Broniatowski et al., 2018). A rumor involves a form of a statement whose veracity is not quickly verifiable or ever confirmed.<sup>10</sup>

**Labels:** To categorize the tweets in terms of their harmfulness, we defined the following labels, again using a Likert scale, where a higher value indicates a higher degree of harm:

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

<sup>9</sup>The use of information as a weapon to spread misinformation and mislead people.

<sup>10</sup><https://en.wikipedia.org/wiki/Rumor>

## Examples:

1. *How convenient but not the least bit surprising from Democrats! As usual they put politics over American citizens. @SpeakerPelosi withheld #coronavirus bill so DCCC could run ads AGAINST GOP candidates! #tcot*

**Label: 5. YES, definitely harmful**

**Explanation:** This tweet is weaponized to target Nancy Pelosi and the Democrats in general.

2. *As we saw over the wkend, disinfo is being spread online about a supposed national lockdown and grounding flights. Be skeptical of rumors. Make sure you're getting info from legitimate sources. The @WhiteHouse is holding daily briefings and @cdcgov is providing the latest.*

**Label: 1. NO, definitely not harmful**

**Explanation:** This tweet is informative and gives advice. It does not attack anyone and is not harmful.

## D.5 Need for Verification

**Question 5:** Do you think that a professional fact-checker should verify the claim in the tweet?

It is important that a verifiable factual check-worthy claim be verified by a professional fact-checker, as the claim may cause harm to society, specific person(s), company(s), product(s), or some government entities. However, not all factual claims are important or worth fact-checking by a professional fact-checker, as this very time-consuming. Therefore, the purpose is to categorize the tweet using the labels defined below. While doing so, the annotator can rely on the answers to the previous questions. For this question, we defined the following labels to categorize the tweets. This question is to be answered, taking the responses to the previous questions into account.

**Labels:**

1. **NO, no need to check:** the tweet does not need to be fact-checked, e.g., because it is not interesting, a joke, or does not contain any claim.
2. **NO, too trivial to check:** the tweet is worth fact-checking, however, this

does not require a professional fact-checker, i.e., a non-expert might be able to fact-check the claim. For example, one can verify the information using reliable sources such as the official website of the WHO, etc. An example of a claim is as follows: “*The GDP of the USA grew by 50% last year.*”

3. **YES, not urgent:** the tweet should be fact-checked by a professional fact-checker, however, this is not urgent or critical;
4. **YES, very urgent:** the tweet can cause immediate harm to a large number of people; therefore, it should be verified as soon as possible by a professional fact-checker;
5. **Not sure:** the content of the tweet does not have enough information to make a judgment. It is recommended to categorize the tweet using this label when the content of the tweet is not understandable at all. For example, it uses a language (i.e., non-English) or references that it is difficult to understand.

#### Examples:

1. *Things the GOP has done during the Covid-19 outbreak: - Illegally traded stocks - Called it a hoax - Blamed it on China - Tried to bailout big business without conditions What they haven't done: - Help workers - Help small businesses - Produced enough tests or ventilators*  
**Label: 2. YES, very urgent**  
**Explanation:** The tweet blames the authorities, and thus, it is important to verify it quickly by a professional fact-checker. In addition, the attention of government entities might be required in order to take necessary actions.
2. **ALERT !!!!!** *The corona virus can be spread through internationally printed albums. If you have any albums at home, put on some gloves, put all the albums in a box and put it outside the front door tonight. I'm collecting all the boxes tonight for safety. Think of your health.*  
**Label: 5. NO, no need to check**  
**Explanation:** This is clearly a joke, and thus it does not require to be checked by a professional fact-checker.

## D.6 Harmful to Society

### Question 6: Is the tweet harmful to society and why?

This question asks whether the content of the tweet is intended to harm or is weaponized to mislead the society. To identify that, we defined the following labels for the categorization.

#### Labels:

- A. **NO, not harmful:** the content of the tweet would not harm the society (e.g., “*I like corona beer*”).
- B. **NO, joke or sarcasm:** the tweet contains a joke (e.g., “*If Corona enters Spain, it'll enter from the side of Barcelona defense*”) or sarcasm (e.g., “*‘The corona virus is a real thing.’ – Wow, I had no idea!*”).
- C. **Not sure:** if the content of the tweet is not understandable enough to judge.
- D. **YES, panic:** the tweet spreads panic. The content of the tweet can cause sudden fear and anxiety for a large part of the society (e.g., “*there are 50,000 cases of COVID-19 in Qatar*”).
- E. **YES, xenophobic, racist, prejudices, or hate-speech:** the tweet spreads xenophobia, racism, or prejudices. According to the dictionary<sup>11</sup> *Xenophobic* refers to fear or hatred of foreigners, people from different cultures, or strangers. *Racism* is the belief that groups of humans possess different behavioral traits corresponding to physical appearance and can be divided based on the superiority of one race over another.<sup>12</sup> It may also refer to prejudice, discrimination, or antagonism directed against other people because they are of a different race or ethnicity. *Prejudice* is an unjustified or incorrect attitude (i.e., typically negative) towards an individual based solely on the individual's membership in a social group.<sup>13</sup> Here is an example: “*do not buy cucumbers from Iran*”.
- F. **YES, bad cure:** the tweet promotes questionable cure, medicine, vaccine, or prevention procedures (e.g., “*... drinking bleach can help cure coronavirus*”).

<sup>11</sup><https://www.dictionary.com/>

<sup>12</sup><https://en.wikipedia.org/wiki/Racism>

<sup>13</sup><http://www.simplypsychology.org/prejudice.html>

G. **YES, rumor, or conspiracy:** the tweet spreads rumors. Rumor is defined as a “specific (or topical) proposition for belief passed along from person to person usually by word of mouth without secure standards of evidence being present” (Allport and Postman, 1947). For example, “*BREAKING: Trump could still own stock in a company that, according to the CDC, will play a major role in providing coronavirus test kits to the federal government, which means that Trump could profit from coronavirus testing. #COVID-19 #coronavirus <https://t.co/Kwl3ylMZRk>*”

H. **YES, other:** if the content of the tweet does not belong to any of the above categories, then this category can be chosen to label the tweet.

#### D.7 Requires Attention

**Question 7:** Do you think that this tweet should get the attention of policy makers of government entities?

Most often people tweet by blaming authorities, providing advice, and/or call for action. Policy makers might want to respond or to react to this. The purpose of this question is to categorize such information. It is important to note that not all information requires attention from a government entity. Therefore, even if the tweet’s content belongs to any of the positive categories, it is important to understand whether it requires attention. For the annotation, it is mandatory to first decide whether attention is necessary (i.e., **YES/NO**). If the answer is **YES**, it is obligatory to select a category from the **YES** sub-categories below.

**Labels:**

A. **NO, not interesting:** the content of the tweet is not important or interesting for any government entity to pay attention to.

B. **Not sure:** if the content of the tweet is not understandable enough to judge;

C. **YES, categorized as in question 6:** some government entity needs to pay attention to this tweet as it is harmful for society and it was labeled as any of the **YES** sub-categories in question 6;

D. **YES, other:** if the tweet cannot be labeled as any of the above categories, then this label should be selected;

E. **YES, blames authorities:** the tweet blames authorities, e.g., “*Dear @VP Pence: Is the below true? Do you have a plan? Also, when are local jurisdictions going to get the #Coronavirus test kits you promised?*”;

F. **YES, contains advice:** the tweet contains advice about social, political, national, or international issues that requires attention from some government entity (e.g., *The elderly & people with pre-existing health conditions are more susceptible to #COVID19. To stay safe, they should: ✓ Keep distance from people who are sick ✓ Frequently wash hands with soap & water ✓ Protect their mental health*);

G. **YES, calls for action:** the tweet states that some government entity should take action for a particular issue (e.g., *I think the Government should close all the Barber Shops and Salons, let people buy shaving machines and other beauty gadgets keep in their houses. Salons and Barbershops might prove to be another Virus spreading channels @citizentvkenya @SenMutula @CSMu-tahi\_Kagwe*);

H. **YES, discusses action taken:** the tweet discusses actions taken by governments, companies, individuals for any particular issue, for example, closure of bars, conferences, churches due to the corona virus (e.g., *Due to the current circumstances with the Corona virus, The 4th Mediterranean Heat Treatment and Surface Engineering Conference in Istanbul postponed to 26-28 Mays 2021.*).

I. **YES, discusses cure:** attention is needed by some government entity as the tweet discusses a possible cure, vaccine, or treatment for a disease;

J. **YES, asks question:** the tweet asks a question about a particular issue and it requires attention from government entities (e.g., *Special thanks to all doctors and nurses, new found respect for you’ll. Is the virus going to totally disappear in the summer? I live in USA and praying that when the temperature warms up the virus will go away...is my thinking accurate?*)

## E Annotation Agreement

In Tables 14, 15, 16 and 17, we report the inter-annotator agreement for English,<sup>14</sup> Arabic, Bulgarian and Dutch, respectively. Overall, we can see that there is moderate to substantial agreement for all questions both for the binary and for the multi-label setting.

Agree. Pair	Q1	Q2	Q3	Q4	Q5	Q6	Q7
<b>Multiclass</b>							
A1 - C	0.81	0.73	0.59	0.74	0.79	0.67	0.73
A2 - C	0.67	0.53	0.44	0.45	0.39	0.65	0.47
A3 - C	0.78	0.58	0.63	0.61	0.70	0.17	0.42
<b>Avg</b>	<b>0.75</b>	<b>0.61</b>	<b>0.55</b>	<b>0.60</b>	<b>0.63</b>	<b>0.50</b>	<b>0.54</b>
<b>Binary</b>							
A1 - C	0.81	0.73	0.77	0.85	0.84	0.77	0.92
A2 - C	0.67	0.58	0.53	0.43	0.52	0.33	0.57
A3 - C	0.78	0.70	0.63	0.70	0.74	0.11	0.57
<b>Avg</b>	<b>0.75</b>	<b>0.67</b>	<b>0.64</b>	<b>0.66</b>	<b>0.70</b>	<b>0.40</b>	<b>0.69</b>

Table 14: Inter-annotator agreement using Fleiss Kappa ( $\kappa$ ) for the **English** dataset. *A* refers to annotator, and *C* refers to consolidation.

Agree. Pair	Q1	Q2	Q3	Q4	Q5	Q6	Q7
<b>Multiclass</b>							
A1 - C	0.58	0.5	0.52	0.53	0.4	0.61	0.47
A2 - C	0.59	0.52	0.52	0.55	0.44	0.62	0.4
A3 - C	0.57	0.44	0.48	0.37	0.36	0.4	0.3
<b>Avg</b>	<b>0.58</b>	<b>0.49</b>	<b>0.51</b>	<b>0.48</b>	<b>0.4</b>	<b>0.54</b>	<b>0.39</b>
<b>Binary</b>							
A1 - C	0.58	0.52	0.53	0.58	0.47	0.65	0.45
A2 - C	0.59	0.57	0.57	0.59	0.47	0.67	0.36
A3 - C	0.57	0.48	0.53	0.47	0.39	0.46	0.29
<b>Avg</b>	<b>0.58</b>	<b>0.52</b>	<b>0.54</b>	<b>0.55</b>	<b>0.44</b>	<b>0.59</b>	<b>0.37</b>

Table 15: Inter-annotator agreement using Fleiss Kappa ( $\kappa$ ) for the **Arabic** dataset. *A* refers to annotator, and *C* refers to consolidation.

Agree. Pair	Q1	Q2	Q3	Q4	Q5	Q6	Q7
<b>Multiclass</b>							
A1 - C	0.77	0.44	0.64	0.53	0.49	0.53	0.51
A2 - C	0.51	0.40	0.59	0.49	0.44	0.56	0.53
A3 - C	0.47	0.38	0.57	0.49	0.38	0.53	0.40
<b>Avg</b>	<b>0.58</b>	<b>0.41</b>	<b>0.60</b>	<b>0.50</b>	<b>0.44</b>	<b>0.54</b>	<b>0.48</b>
<b>Binary</b>							
A1 - C	0.77	0.41	0.71	0.56	0.61	0.47	0.50
A2 - C	0.51	0.39	0.64	0.52	0.57	0.51	0.53
A3 - C	0.47	0.34	0.62	0.52	0.54	0.47	0.38
<b>Avg</b>	<b>0.58</b>	<b>0.38</b>	<b>0.66</b>	<b>0.53</b>	<b>0.57</b>	<b>0.48</b>	<b>0.47</b>

Table 16: Inter-annotator agreement using Fleiss Kappa ( $\kappa$ ) for the **Bulgarian** dataset. *A* refers to annotator, and *C* refers to consolidation.

Agree. Pair	Q1	Q2	Q3	Q4	Q5	Q6	Q7
<b>Multiclass</b>							
A1 - C	0.63	0.54	0.58	0.58	0.54	0.66	0.63
A2 - C	0.83	0.69	0.68	0.70	0.65	0.59	0.62
A3 - C	0.76	0.64	0.59	0.59	0.62	0.51	0.59
<b>Avg</b>	<b>0.74</b>	<b>0.62</b>	<b>0.62</b>	<b>0.62</b>	<b>0.60</b>	<b>0.59</b>	<b>0.61</b>
<b>Binary</b>							
A1 - C	0.63	0.62	0.63	0.60	0.60	0.68	0.69
A2 - C	0.83	0.73	0.76	0.77	0.75	0.63	0.69
A3 - C	0.76	0.68	0.68	0.66	0.69	0.53	0.65
<b>Avg</b>	<b>0.74</b>	<b>0.67</b>	<b>0.69</b>	<b>0.68</b>	<b>0.68</b>	<b>0.61</b>	<b>0.68</b>

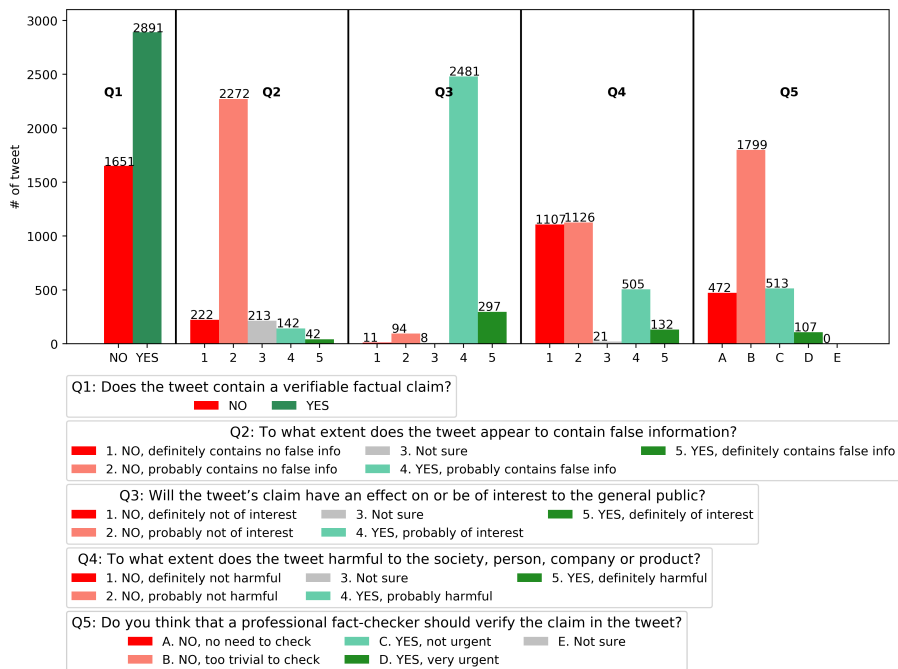
Table 17: Inter-annotator agreement using Fleiss Kappa ( $\kappa$ ) for the **Dutch** dataset. *A* refers to annotator, and *C* refers to consolidation.

<sup>14</sup>We have already presented the table for English in the main text of the paper, but we repeat it here to facilitate cross-language comparisons.

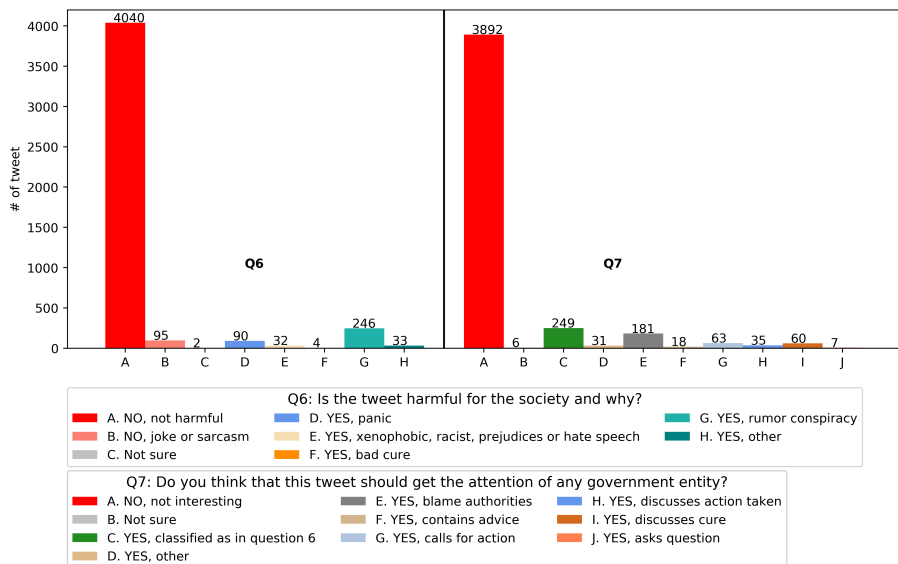


## **F Class Label Distribution**

Figures 3, 4, 5 and 6 report detailed statistics about the label distribution for the manual annotations for each question in English, Arabic, Bulgarian, and Dutch, respectively.

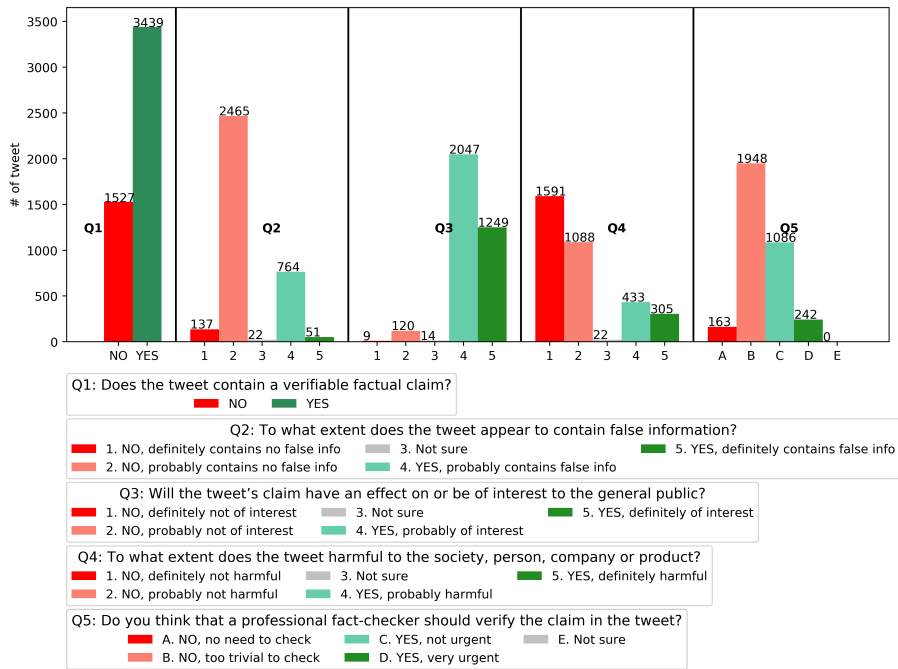


(a) Questions (Q1-5).

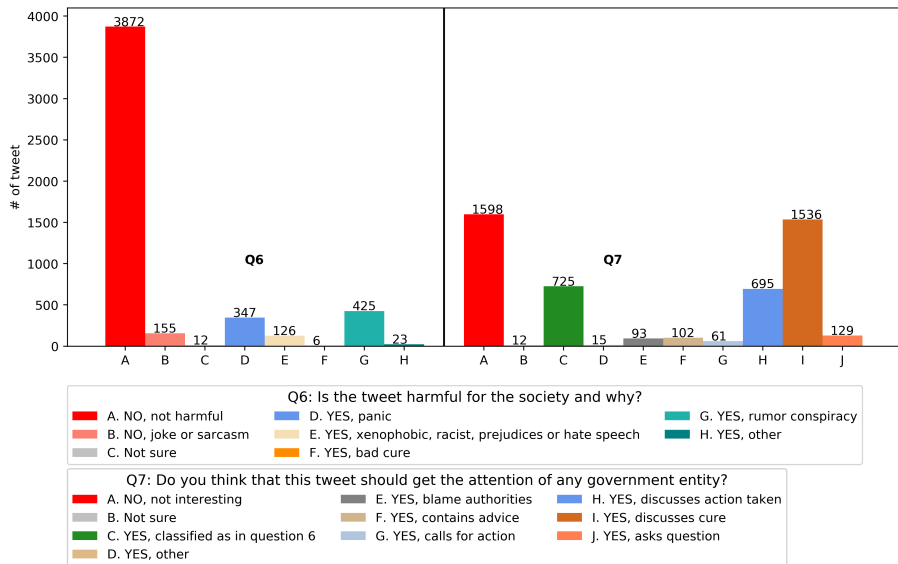


(b) Questions (Q6-7).

Figure 3: Distribution of class labels for **English tweets**

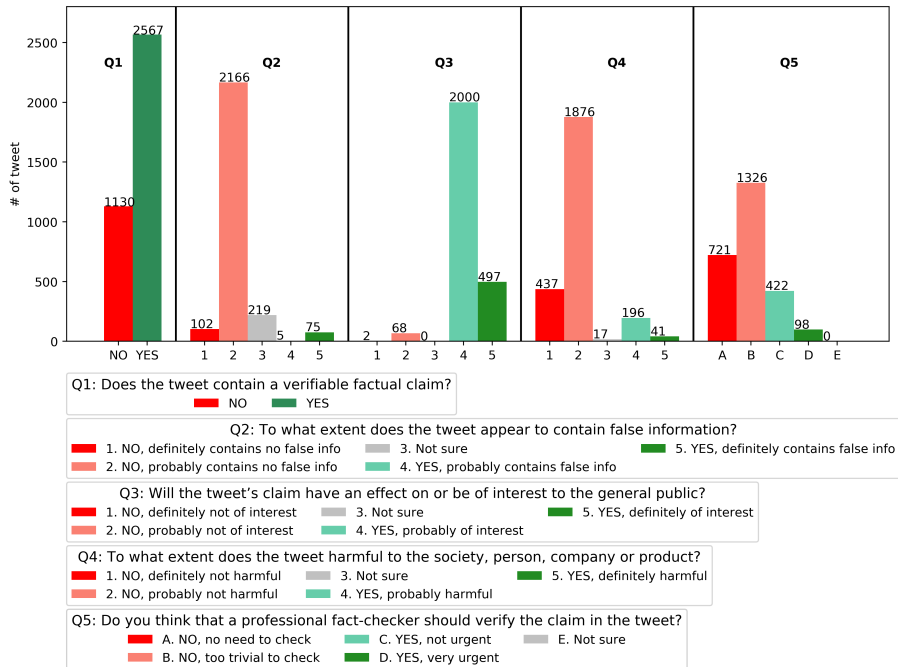


(a) Questions (Q1-5).

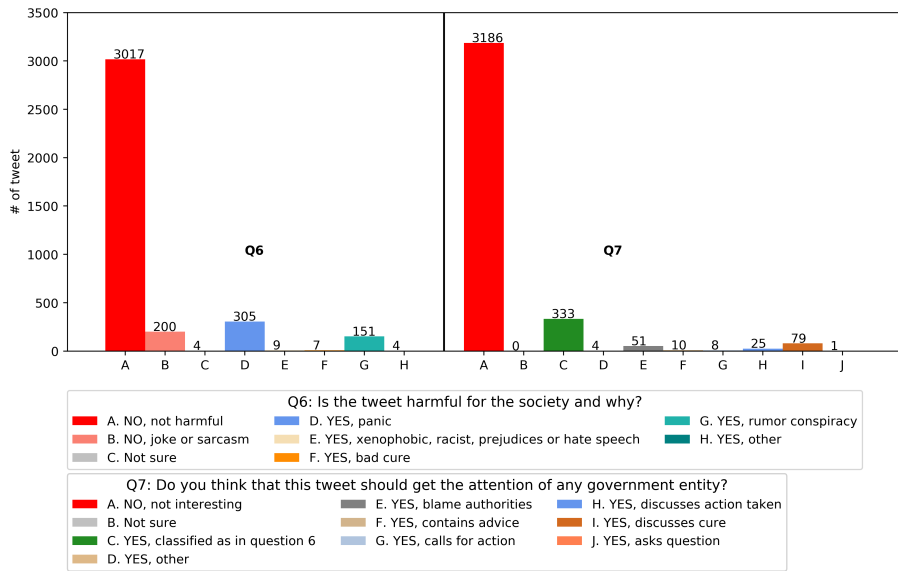


(b) Questions (Q6-7).

Figure 4: Distribution of class labels for Arabic tweets

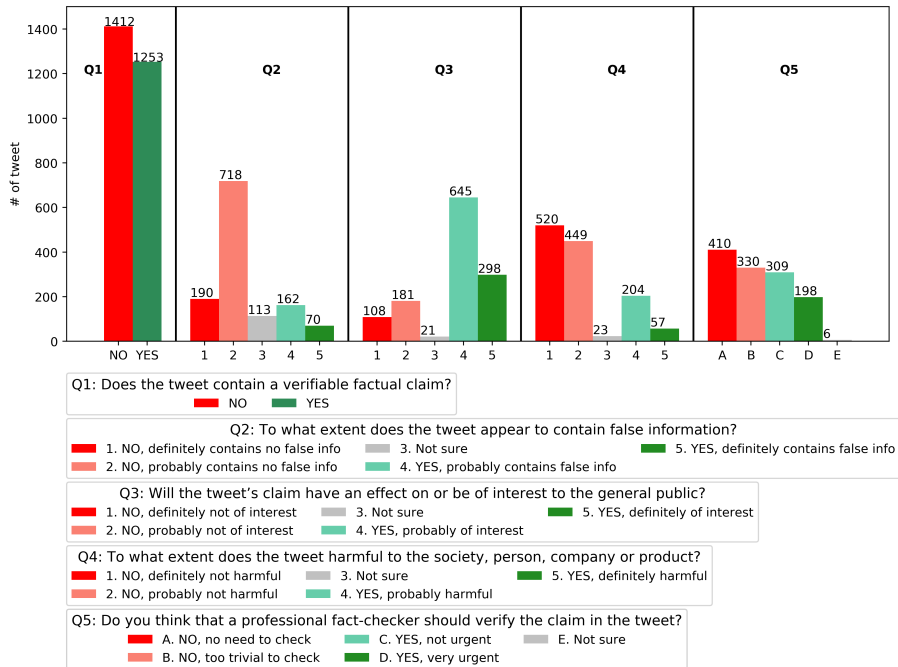


(a) Questions (Q1-5).

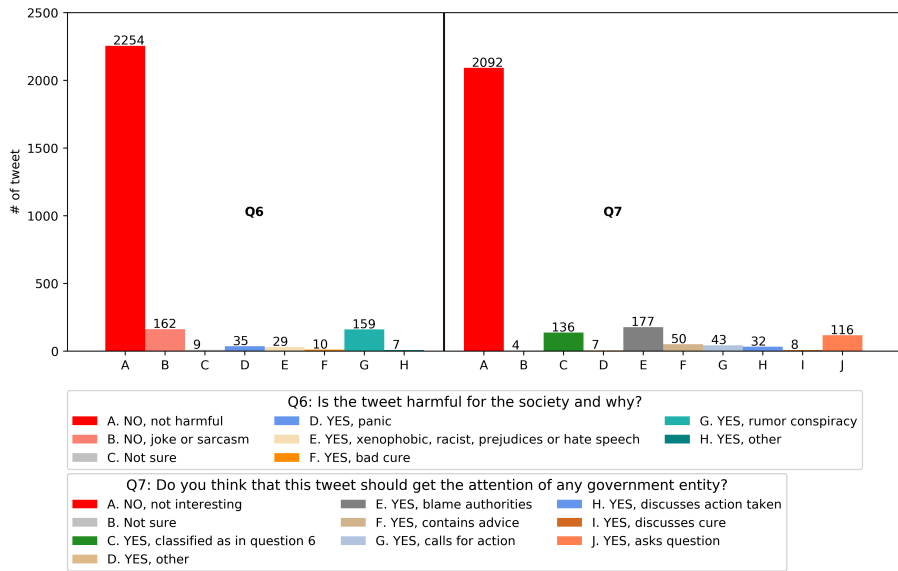


(b) Questions (Q6-7).

Figure 5: Distribution of class labels for **Bulgarian tweets**



(a) Questions (Q1-5).



(b) Questions (Q6-7).

Figure 6: Distribution of class labels for Dutch tweets

## G Correlation Between Questions

Finally, we study the correlation between the labels for different questions across for each of the four languages.

### G.1 English Tweets

Figure 7 shows contingency and correlation tables in a form of a heatmap for different question pairs obtained from the English tweet dataset. For questions Q2-3, it appears that there is a high association between “...no false info” and the general public interest as shown in Figure 7a. For questions Q2 and Q4 (Figure 7b), strong association can be observed between “... no false info” and “... not harmful” (86%) compared to “harmful” (13%) for either an individual, a products or government entities. By analyzing questions Q2 and Q5 (Figure 7c), we conclude that “... no false info” is associated with either “no need to check” or “too trivial to check”, highlighting the fact that a professional fact-checker does not need to spend time on them. From questions Q3 and Q4 (Figure 7d), it appears that when the content of the tweets is “not harmful” the general public interest is higher (74%) than when it is “harmful” (22%). From question Q3 and Q5 (Figure 7e), we see an interesting phenomenon, namely that tweets of high general public interest have a greater association with a professional fact-checker having to verify them (22%) compared to either “too trivial to check” or “no need to check” (78%). Questions Q4 and Q5 (Figure 7f) show that “harmful” tweets require an attention (69%) from a professional fact-checkers than “not harmful” tweets (30%). Our findings for Q6 and Q7 (Figure 7g) suggest that the majority of the tweets are not harmful to society, which also requires less attention from government entities. The third most common tweet label for Q7 blames the authorities, even though they are mostly not harmful for society.

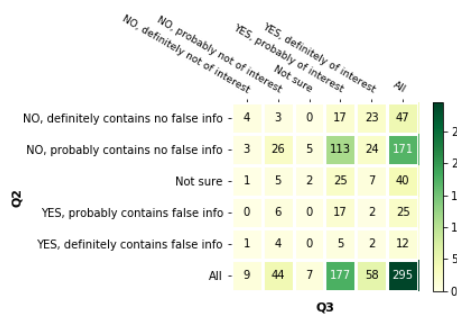
We computed the correlation using the Likert scale values (i.e., 1-5) that we defined for these questions. We observed that overall Q2 and Q3 are negatively correlated, which suggests that if the claim contains no false information, it is of high interest to the general public. This can be also observed in Figure 7a. Questions Q2 and Q4 exhibit positive correlation, which might be due to their high association with “...no false info” and “... not harmful”.

### G.2 Arabic Tweets

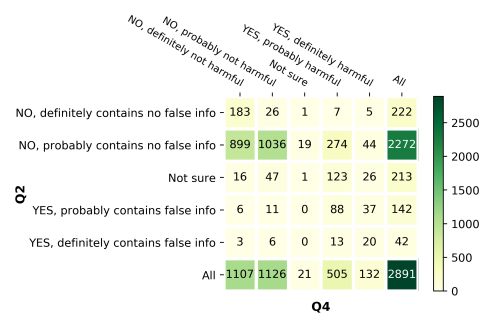
Figure 8 shows similar heatmaps for the Arabic tweets. For Q2 and Q3 (Figure 8a), we can observe that the association between “... contains no false info” and general public interest is higher (76%) than “... contains false info” (23%). From questions Q2 and Q4 (Figure 8b), we can conclude that “...contains no false info” is associated with “...not harmful” and “...contains false info” is associated with “... harmful”. From the relation between Q2 and Q5 (Figure 8c), we can observe that in the majority of the cases “...contains no false info” is associated with either “no need to check” or “too trivial to check”, which means that a professional fact-checker does not need to verify them. The analysis between questions Q3 and Q4 suggests that the general public interest is higher when the content of the tweets is not harmful (79%) than when it is harmful (21%) (Figure 8d). From questions Q3 and Q5, we can observe that the general public interest is higher when the claim(s) in the tweets are either “no need to check” or “too trivial to check” (Figure 8e). The analysis between question Q4 and Q5 shows that “not harmful” tweets are either “no need to check” or “too trivial to check” by a professional fact-checker (Figure 8f). From questions Q6 and Q7, we notice that in the majority of the cases the tweets are not harmful for society and hence they are not interesting for government entities (Figure 8g).

### G.3 Bulgarian and Dutch Tweets

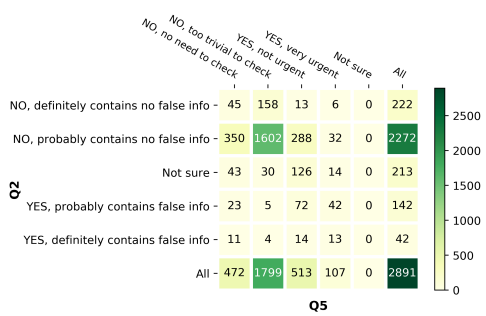
Figures 9 and 10 show the same kinds of heatmaps for the Bulgarian and the Dutch datasets, respectively. The observations are very similar.



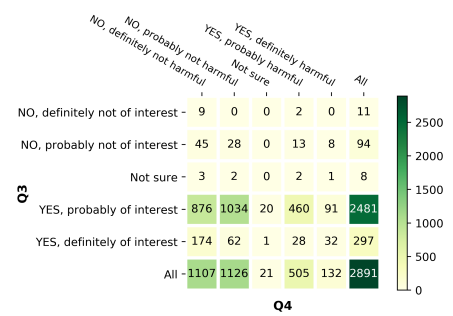
(a) Heatmap for Q2 and Q3.



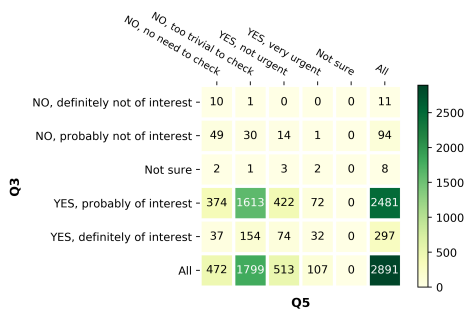
(b) Heatmap for Q2 and Q4.



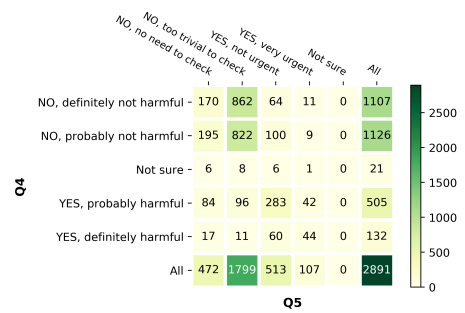
(c) Heatmap for Q2 and Q5.



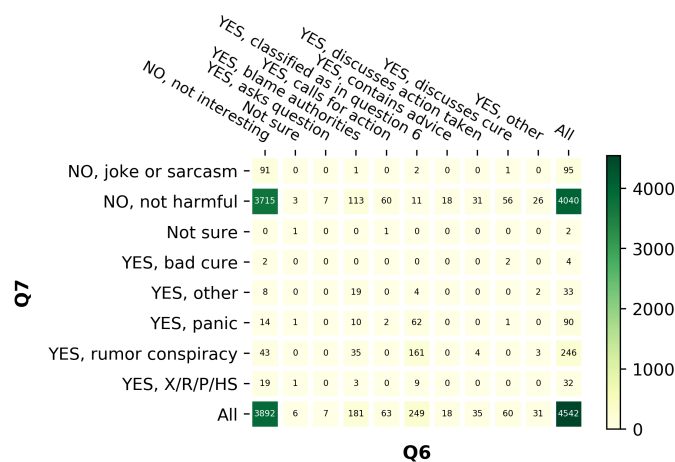
(d) Heatmap for Q3 and Q4.



(e) Heatmap for Q3 and Q5.

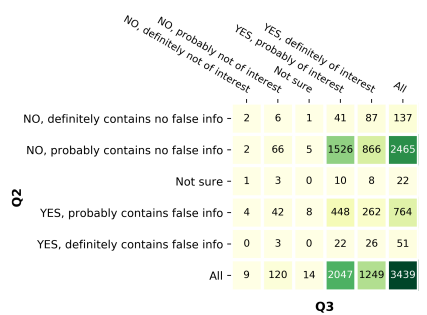


(f) Heatmap for Q4 and Q5.

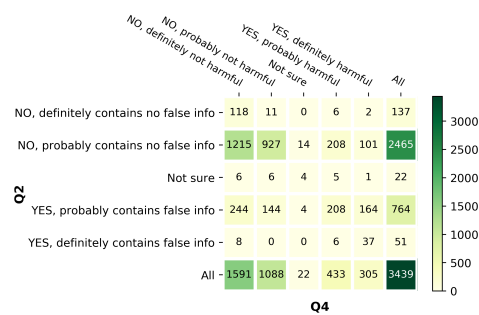


(g) Heatmap for Q6 and Q7. YES, X/R/P/HS – YES, xenophobic, racist, prejudices or hate speech

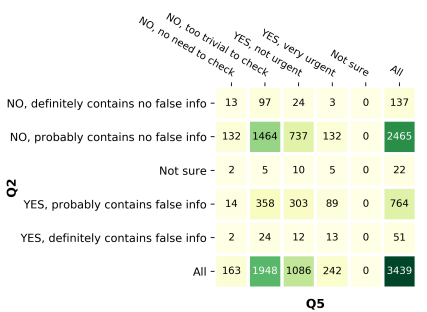
Figure 7: Contingency and correlation heatmaps for the **English tweets** for different question pairs.



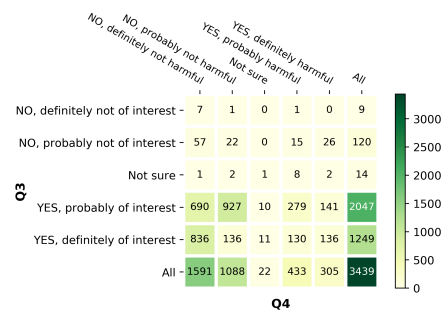
(a) Heatmap for Q2 and Q3.



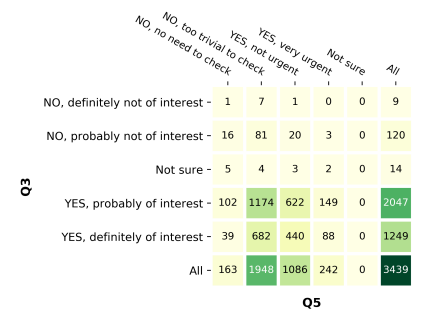
(b) Heatmap for Q2 and Q4.



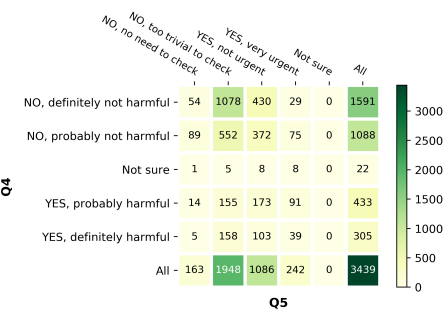
(c) Heatmap for Q2 and Q5.



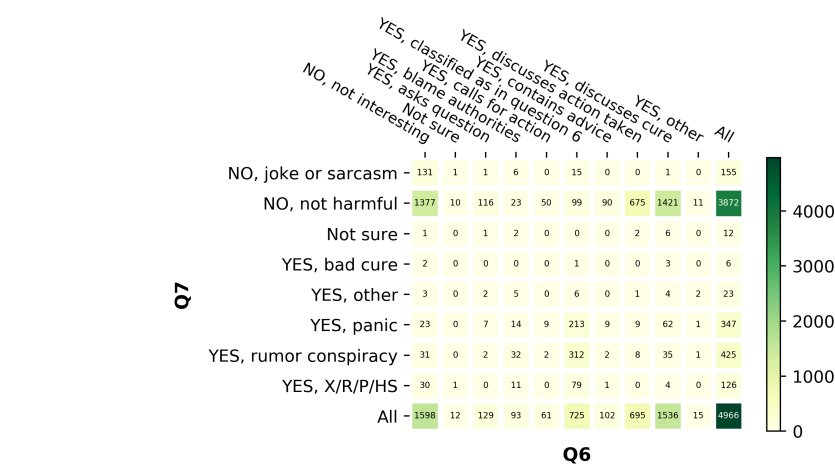
(d) Heatmap for Q3 and Q4.



(e) Heatmap for Q3 and Q5.



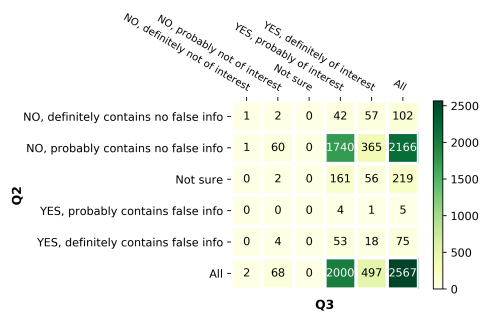
(f) Heatmap for Q4 and Q5.



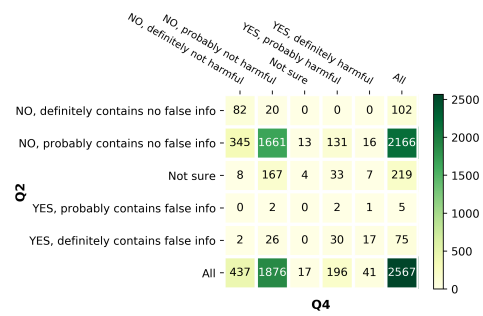
(g) Heatmap for Q6 and Q7. YES, X/R/P/HS – YES, xenophobic, racist, prejudices or hate speech

Figure 8: Contingency and correlation heatmaps for the Arabic tweets for different question pairs.

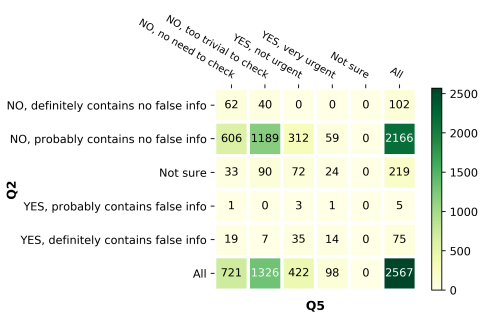




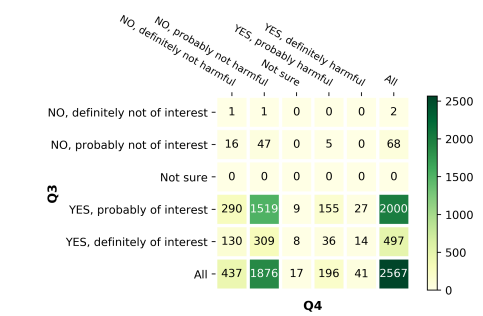
(a) Heatmap for Q2 and Q3.



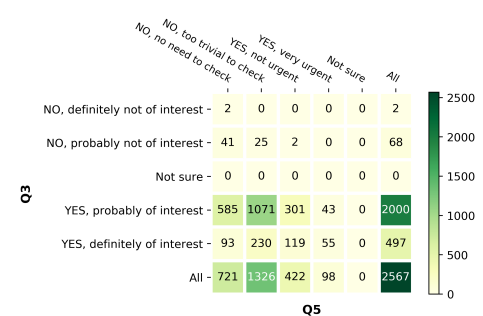
(b) Heatmap for Q2 and Q4.



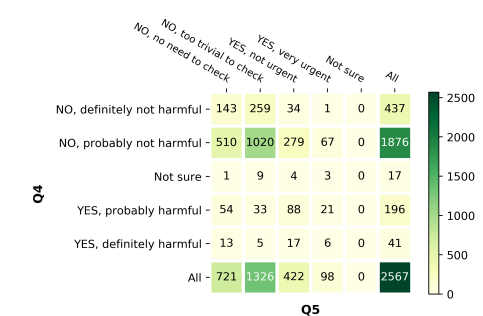
(c) Heatmap for Q2 and Q5.



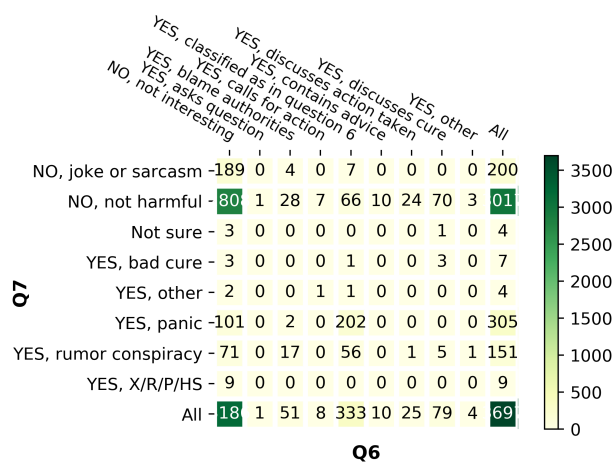
(d) Heatmap for Q3 and Q4.



(e) Heatmap for Q3 and Q5.

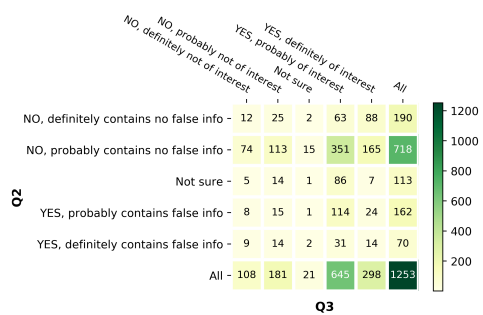


(f) Heatmap for Q4 and Q5.

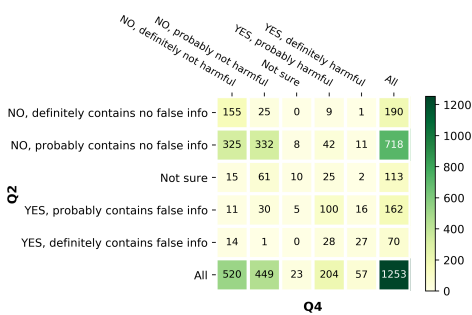


(g) Heatmap for Q6 and Q7. YES, X/R/P/HS – YES, xenophobic, racist, prejudices or hate speech

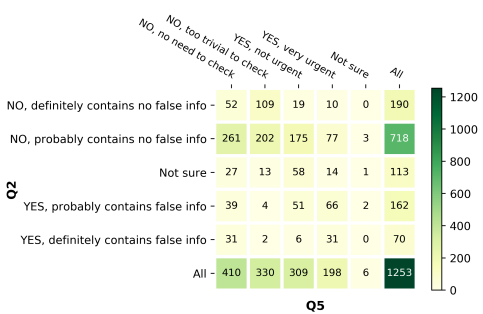
Figure 9: Contingency and correlation heatmaps for the **Bulgarian tweets** for different question pairs.



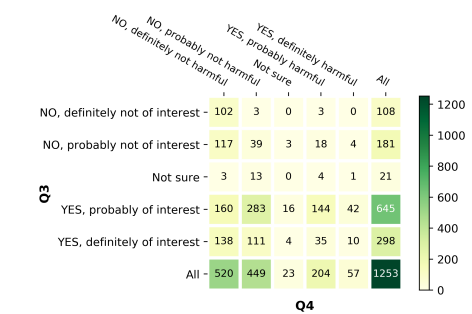
(a) Heatmap for Q2 and Q3.



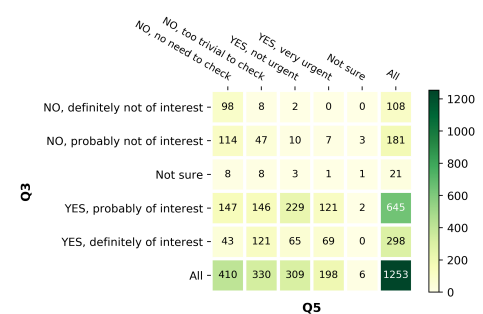
(b) Heatmap for Q2 and Q4.



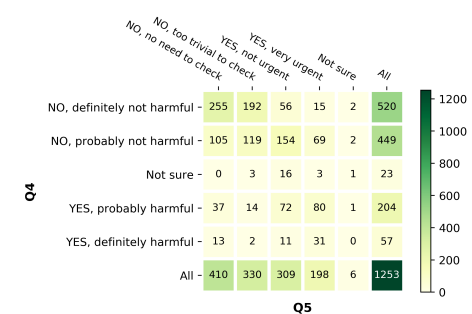
(c) Heatmap for Q2 and Q5.



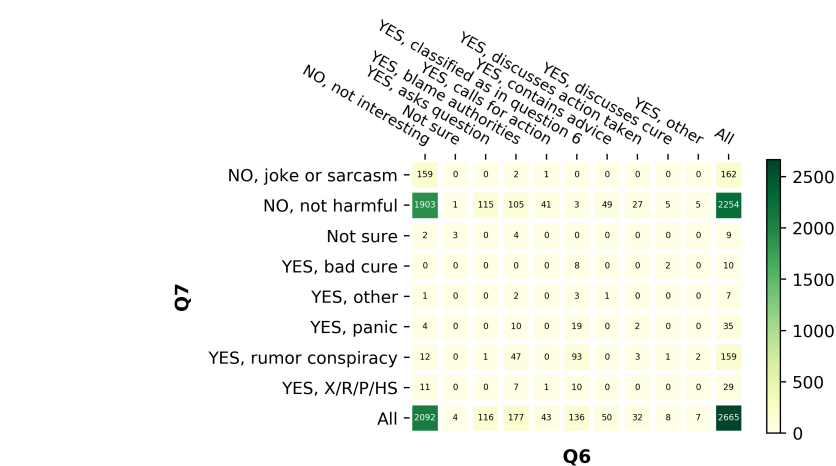
(d) Heatmap for Q3 and Q4.



(e) Heatmap for Q3 and Q5.



(f) Heatmap for Q4 and Q5.



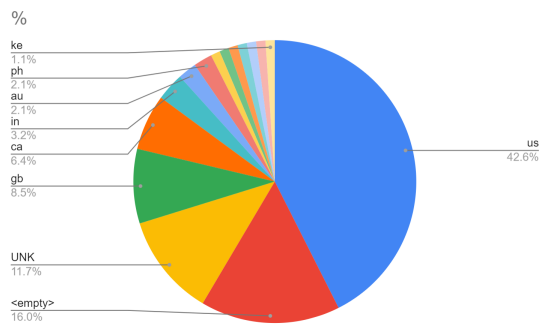
(g) Heatmap for Q6 and Q7. YES, X/R/P/HS – YES, xenophobic, racist, prejudices or hate speech

Figure 10: Contingency and correlation heatmaps for the Dutch tweets for different question pairs.

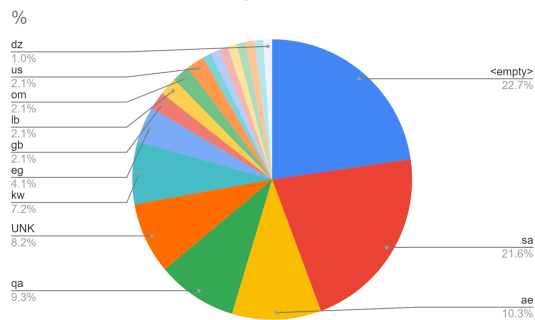
## H Geographical Distribution: English and Arabic

Figure 11 shows the geographical distribution of the annotated tweets for English and Arabic. We consider the country of the tweet author or the original author in case of a retweet. We can see that most English tweets come from USA, UK, Canada, and India, while most Arabic tweets come from the Gulf region (KSA, UAE, Qatar, and Kuwait). Yet, for both languages, we have tweets from multiple countries, which means that there is good diversity of interests, topics, style, etc.

We did not perform this analysis for Bulgarian and Dutch, as these are less international languages, and their speakers are mostly concentrated in Bulgaria and the Netherlands, respectively.



(a) English dataset



(b) Arabic dataset

Figure 11: Distribution by country for English and Arabic tweets.

## **I Verified and Unverified Accounts: English and Arabic**

We study the correlation between tweet labels and whether or not the original author of a tweet has a verified account. Verified accounts include such by government entities, public figures, celebrities, etc., which have a large number of followers, and thus their tweets typically have higher impact.

Figure 12 shows that verified accounts tend to post more tweets that contain factual claims than unverified accounts (Q1), and their tweets are less likely to contain false information (Q2), are more likely to be of interest to the general public (Q3), and are less likely to be harmful (Q4 and Q6).

Based on this study, we have added corresponding features for our models.

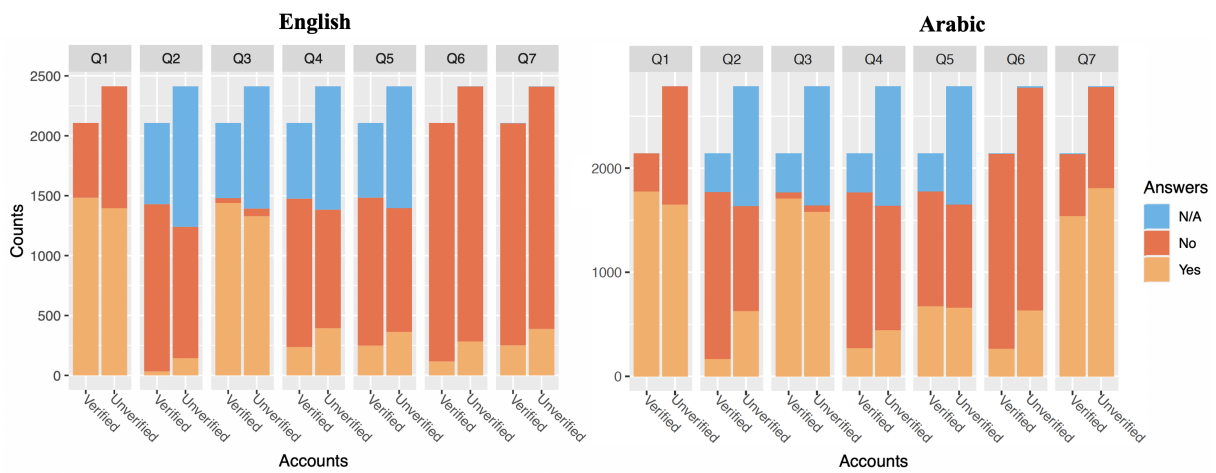


Figure 12: Verified vs. non-verified account distribution for English and Arabic across the different questions. NA refers to tweets that have not been labeled for those questions, they are identical to the tweets categorized with the label NO in Q1.

## **J Multimedia in Tweets: English and Arabic**

In this subsection, we study the correlation between whether a tweet contains multimedia (image or video) and the annotation labels. Generally, people trust videos more than images or plain texts, which suggests that tweets with video could have a higher impact. Figure 13 shows the distribution of media types for English and Arabic.

We can see that if a tweet contains multimedia content, it is likely to contain a factual claim (Q1), to have a higher impact to the general public (Q3), but it is less likely to contain false information (Q2) or to be harmful to the society (Q4). These observations in part motivated us to model the use of multimedia as part of our features.

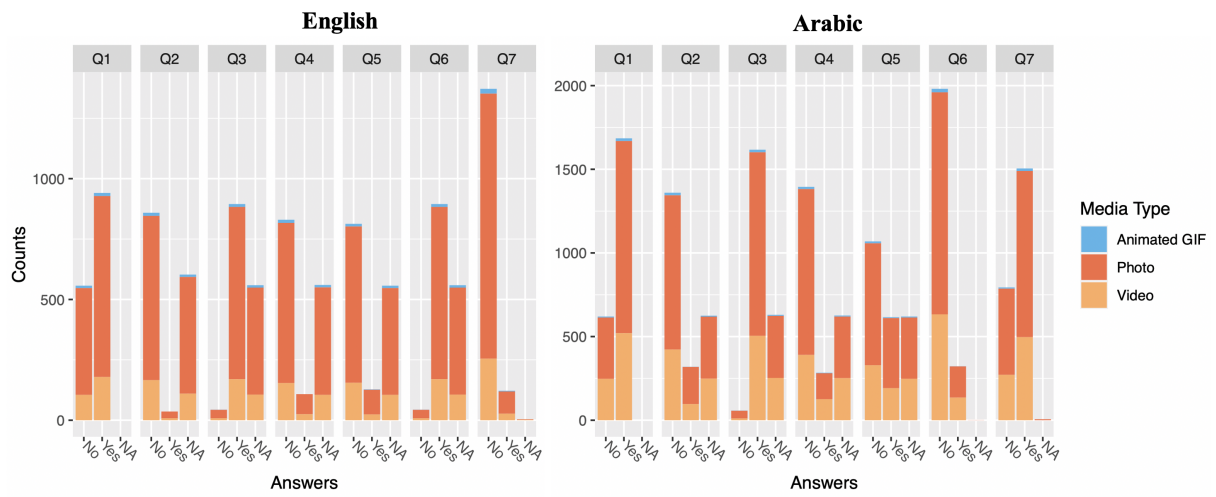


Figure 13: Distribution of media types in English and Arabic tweets.