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

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov*, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Preslav Nakov
Qatar Computing Research Institute, HBKU, Qatar
*Sofia University, Sofia, Bulgaria

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 is ranked second in the list 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. Thus, here we design, annotate, and release to the research community a new dataset 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 as a whole, and (*iii*) covers both English and Arabic. Finally, we show strong evaluation results using state-of-the-art Transformers, thus confirming the practical utility of the annotation schema and of the dataset.

1 Introduction

The rise of social media has made them one of the main communication channels for information dissemination and consumption. In fact, many people rely on social media as their primary source of news instead of traditional news (Perrin, 2015). Unfortunately, the democratic nature of social media, where anybody can easily become a news producer, has raised questions about the quality and the factuality of the information that is shared on these platforms. Eventually, social media have become the main channel to spread disinformation (Kumar and Shah, 2018; Alzanin and Azmi, 2018).

Figure 1 shows some example tweets that demonstrate how online users discuss topics related to COVID-19 in social media. As we can see, the problem is much broader than simply looking at factuality, and the tweets we show could be of interest to journalists, fact-checkers, social media platforms, policymakers, and society in general. The examples include tweets spreading rumors (Figure 1a), promoting conspiracy theories (Figure 1h), making jokes (Figure 1c), instilling panic (Figure 1b), promoting fake cures (Figure 1d), or spreading xenophobia, racism, and prejudices (Figure 1e). Other examples contain information that could be potentially useful and might deserve the attention and some action/reaction by government entities. For example, the tweet in Figure 1f blames the authorities for their inaction regarding COVID-19 testing. The tweets in the Figures 1f and 1g are also useful both for authorities and for the general public as they discuss actions a government has taken to fight the pandemic and suggests actions that probably should be taken.

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.

There has been a lot of fact-checking effort by about 200 organizations worldwide,[1] such as PolitiFact, FactCheck, Snopes, and Full Fact, but it is all manual and very time-consuming, and thus it often comes too late and it does not scale. Moreover, the focus has been exclusively on factuality and to some extent on check-worthiness; yet, when it comes to COVID-19, there is a broader range of relevant issues, as the above examples show.

[1] http://tiny.cc/zd1fnz

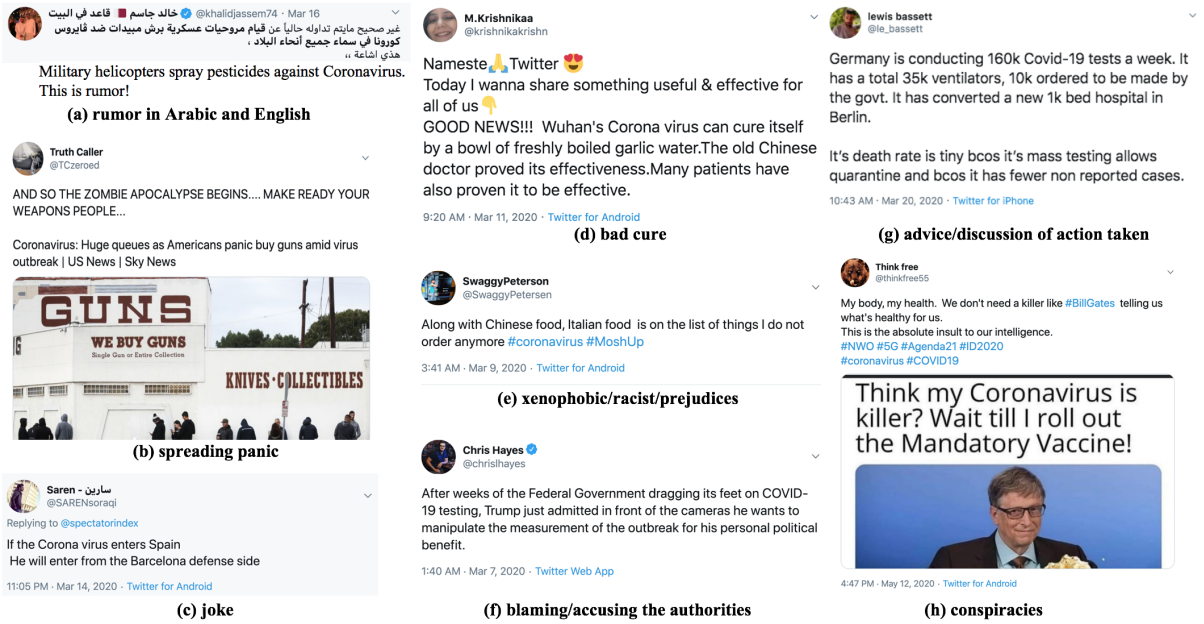


Figure 1: Example tweets, which would be of potential interest to journalists, fact-checkers, social media platforms, policy makers, government entities, and the society as a whole.

Addressing any of the above issues requires significant efforts in terms of (*i*) defining comprehensive annotation guidelines, (*ii*) collecting tweets about COVID-19 and sampling from them for annotation, (*iii*) manually annotating the tweets, and (*iv*) training and evaluating models. Given the interconnected nature of the issues, it is more efficient to address them simultaneously. With this consideration in mind, we adopt a *multifaceted* approach, focusing on three key aspects: (*i*) Does the tweet contain a claim that is worth fact-checking? (*ii*) Is the tweet harmful to the society? and (*iii*) Should a government entity take notice of it? We define seven questions to cover these aspects.

Our contributions are as follows:

- We develop comprehensive guidelines to annotate social media messages that combine the perspectives and the interests of journalists, fact-checkers, social media platforms, policymakers, and the society as a whole.
- We develop an annotated dataset of tweets covering two languages: English and Arabic.
- We present experimental results showing very sizable improvements over the baselines when using coarse-grained yes/no labels. For most questions, we further demonstrate sizable improvements in a fine-grained multiclass setup (5-10 classes per question). Finally, we show the potential for cross-language learning.

2 Related Work

Journalists, online users, and researchers are well-aware of the proliferation of false information, and thus topics such as credibility and fact-checking have become important research topics. The interested reader can learn more about "fake news" from the overview by Shu et al. (2017), which adopted a data mining perspective and focused on social media. Another survey by Thorne and Vlachos (2018) took a fact-checking perspective on "fake news" and related problems. Yet another survey was performed by Li et al. (2016), covering truth discovery in general. Finally, two articles in *Science* offer a general overview and discussion on the science of "fake news" (Lazer et al., 2018) and of the process of proliferation of true and false news online (Vosoughi et al., 2018).

2.1 Fact-Checking

Research on fact-checking claims was 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), *ClaimsKG* dataset and system (Tchechmedjiev et al., 2019) of 28K claims from 8 fact-checking organizations, the *MultiFC* dataset of 38K claims from 26 fact-checking organizations (Augenstein et al., 2019), and the 10K claims *Truth of Various Shades* (Rashkin et al., 2017) dataset, among other smaller-size ones.

There have been also datasets for other languages, created in a similar fashion, e.g., Baly et al. (2018) created a dataset of 402 Arabic claims extracted from Verify-SY. A number of datasets were created as part of shared tasks. In most cases, they did not rely on fact-checking websites, but performed their own annotation, either (a) manually, e.g., the SemEval-2017 task 8 (Derczynski et al., 2017) and the SemEval-2019 task 7 (Gorrell et al., 2019) on Determining Rumour Veracity and Support for Rumours (RumourEval), the SemEval-2019 task 8 on Fact-Checking in Community Question Answering Forums (Mihaylova et al., 2019), the CLEF-2019 on Automatic Identification and Verification of Claims (Elsayed et al., 2019) and the CLEF-2020 Lab on Enabling Automatic Identification and Verification of Claims in Social Media (Barrón-Cedeño et al., 2020), which featured both English and Arabic, or (b) using crowdsourcing, e.g., the FEVER 2018 and 2019 tasks on Fact Extraction and VERification, which focused on fact-checking made-up claims about content present in Wikipedia (Thorne et al., 2018, 2019).

Unlike our work, the above datasets did not focus specifically on tweets (they included claims originating in the news, speeches, political debates, community question answering forums, or just made up by human annotators; RumourEval is a notable exception), targeted factuality only (while we cover a number of other issues), were limited to a single language (typically English; except for the CLEF-2019 and CLEF-2020 labs, which support English and Arabic), and did not focus on COVID-19.

2.2 Check-Worthiness Estimation

Another relevant line of research is on check-worthiness estimation, i.e., detecting claims that should be prioritized for fact-checking. While early work used manual annotations (Hassan et al., 2015), most datasets were once again derived from fact-checking websites, e.g., by observing which claims in a political debate or speech were selected for manual fact-checking (Gencheva et al., 2017; Patwari et al., 2017). Such datasets often originated or were extended as part of shared tasks such as the 2018-2020 editions of the above-mentioned CLEF CheckThat! lab (Nakov et al., 2018; Elsayed et al., 2019; Barrón-Cedeño et al., 2020). These datasets are subject to similar limitations as for the fact-checking datasets above, the most important one being that they focus on a single task.

2.3 COVID-19 Research

In this study, we mainly focus on social media content. We specifically target disinformation posted on Twitter that is related to the COVID-19 pandemic. A recent effort related to the pandemic used the Sina Weibo microblogging platform to study different situational information types, e.g., "caution and advice" (Li et al., 2020). In another study, the authors reported media bias and rumor amplification patterns for COVID-19 (Cinelli et al., 2020) using five different social media platforms. Medford et al. (2020) analyzed COVID-19 related tweets to understand different content types such as emotional, racially prejudiced, xenophobic or content that causes fear.

Other recent work includes identifying low-credibility information using data from social media (Yang et al., 2020), detecting prejudice (Vidgen et al., 2020), finding challenges related to data, tools, and ethical issues (Ding et al., 2020), analyzing the spread of COVID-19 misinformation in relation to culture, society, and politics (Leng et al., 2020), detecting the spread of misleading information and the credibility of users who propagate it (Mourad et al., 2020), identifying positive influencers to propagate information to (Pastor-Escuredo and Tarazona, 2020), analyzing the users who spread misinformation and the propagation of misinformation (Shahi et al., 2020), analyzing psychometric aspects in relation to the COVID-19 infodemic (Jolly et al., 2020), developing a multilingual COVID-19 Instagram dataset (Zarei et al., 2020), and detecting disinformation campaigns (Vargas et al., 2020). All this work focused on a limited set of issues, while here we model the perspectives of journalists, fact-checkers, social media platforms, policy makers, and the society.

3 Dataset

3.1 Data Collection

For this task, we collected COVID-19 related tweets using twarc,[2] a Python wrapper for the Twitter Streaming API. Specifically, we collected tweets that matched a set of COVID-19 related keywords (See Appendix A). We ran queries in two time frames, namely: March 9–10, 2020 and March 20–25, 2020. We filtered out all non-Arabic or non-English tweets, and we annotated the most frequent tweets in English and Arabic.

[2] https://github.com/DocNow/twarc

Tweet 1: *So, the last week I have been battling COVID-19 & Pneumonia. Never in my life have I been this ill. "Young people arent at risk, theyll only have mild symptoms" Wrong. I want to open up about the difficulties Ive gone through these past days, what it was like in the ICU...*		
Q1	YES	Expl: This has a factual claim, in which user posted his personal testimony, mentioning his experience as a COVID-19 patient.
Q2	NO, probably contains no false info	Expl: As the twitter user himself is providing his testimony, therefore, it might be correct information. In addition, the user is a verified user, which makes us to believe that it has a less chance of misinformation.
Q3	YES, probably of interest	Expl: General population might get interest in this how it is like to be a COVID-19 patient.
Q4	NO, probably not harmful	Expl: As it would not harm to anyone, therefore it is not harmful.
Q5	YES, not urgent	Expl: It is a factual claim and worthwhile to fact-check, however, it is less important for the fact-checker.
Q6	NO, not harmful	Expl: It is not harmful for the society as it does not express anything that can affect society.
Q7	YES, blame authorities	Expl: Upon reading the whole threads it seems that user explicitly blames authorities by mentioning "*. . . The government has failed us. Im lucky, others wont be. Its far past the time to take action. Not words, ACTION. Step the fuck up, and protect the people of this country. If they wont, we need to. Stay inside, be smart. No death is worth you being ignorant. We can do this.*".
Tweet 2: *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	Expl: This tweet contains factual claim. This is correlation and causation. The claim is "UK closed the borders because of the Macron's threat".
Q2	NO, probably contains no false info	Expl: It may not contains false info as it came from an authentic person.
Q3	YES, probably of interest	Expl: Many people might be interested for the information in this tweet as the Prime minister took some action to prevent COVID-19.
Q4	YES, definitely harmful	Expl: It is harmful as it blames government officials.
Q5	YES, very urgent	Expl: Professional fact-cheker should verify this immediately as it is attacking government officials.
Q6	YES, rumor, or conspiracy	Expl: The content of the tweet cannot be easily verified as it could be a political move to attack Boris.
Q7	YES, blame authorities	Expl: The content of the tweet clearly blemes authority.

Table 1: Examples of annotated tweets, their labels, and some explanations.

3.2 Annotation Task

The annotation task aims to determine 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 requires verification, and how interesting it is for a government entity to pay attention to. To address them, we defined three goals, namely: (*i*) if there is a claim in the tweet, and is it worth fact-checking? (Q1-5) (*ii*) is the tweet harmful to the society? (Q6) and (*iii*) should a government body or policy makers take notice of it (Q7)? These are then formulated into seven questions presented in Table 2. A complete listing of annotation instructions, with examples, is available in Appendix B.

Although some of the questions are correlated, the annotation instructions are designed so that the dataset can be used independently for different tasks. Questions 2-4 (see Table 2) are designed as both categorical and numerical (i.e., using a Likert scale) in order to enable their use in either classification or regression tasks.

In Table 1, we present a sample tweet, annotated for all questions. *Tweet 1* negates the claim that *"Young people aren't at risk"* through personal testimony of the experience of being a COVID-19 patient. Thus, Question 1 is marked as *Yes*. The tweet probably contains no false information as a verified user is providing information about himself. The tweet is of interest to the general population as it clears the misconception about *"young people not at risk"*. The tweet is not harmful to society but it blames the authorities. *Tweet 2* contains a potentially harmful claim with a causal argument possibly requiring urgent fact-checking. Further, it attacks government officials, which may warrant a response, clarification, or attention from policy makers.

Our comprehensive guidelines can serve as an annotation standard to encourage community efforts towards sorting information based on authenticity and relevance to journalists, fact-checkers, and policymakers. The diversity of annotations enables interesting modeling solutions. Developing such guidelines and the dataset have been challenging due to the concise and noisy nature of messages in social media, some temporal aspects, and the high degree of subjectivity. The temporal aspect refers to the time the tweet was posted, e.g., at time t a claim may be true, but not at time $t + 1$. This time t can be a day, a week, or a month. The guidelines were developed over multiple labeling iterations of examples, and the annotations and guidelines were refined in consolidation meetings. From a modeling perspective, though each question serves as an independent task, some of questions are directly connected and can be considered in relation to each other. For example, the fifth question can be analyzed in relation to the first four questions.

Similarly, all tasks can be combined in a multi-task setting for building one model that serves all the above-described purposes. Another interesting research frontier to explore is how to integrate in the modeling process additional information, such as images, videos, emoticons, or links to external websites that users post as part of their tweets to support their claims. Note that the annotations were carried out while taking into account this supplementary information, even when the tweets were posted as a reply.

3.3 Annotation Challenges

As social media data is noisy and the annotation tasks are highly subjective, disagreement is a typical scenario. The disputed labels were resolved in a consensus meeting. Such an approach has also been used in similar work (Zubiaga et al., 2015). In the cases where disagreements were not resolved among the annotators' group working on the tweets, another consensus meeting was carried out among all the annotators who worked on defining the labels and on improving the annotation guidelines. The annotation task was also time-consuming. For example, on average, an hour was needed for two annotators to resolve the disagreement for 20 tweets.

3.4 Labels for Classification Tasks

The annotation has been designed in a way that fine-grained (i.e., multiclass) labels can be easily transformed into coarse-grained (i.e., binary, Yes *vs* No) labels. Therefore, we transformed multiclass labels for Q2-7 into binary labels considering all ***Yes**** into **Yes**, and all ***No**** into **No** while dropping *not sure* tweets, as reported in Table 2. We used these two sets of labels for two types of experiments: binary and multiclass.

3.5 Data Statistics

Based on the annotation instructions, in the first phase, we annotated 504 tweets in English and 218 in Arabic.[3] Each tweet have been judged by at least three annotators. In total, seven annotators were involved in the process. In Table 2 we report the distribution of class labels of the annotated English and Arabic tweets. Though we focus the following analysis on English tweets, the distribution of

[3]Note that our annotation task is currently ongoing and we expect to annotate more tweets in the near future. We will make them available on ANONYMOUS

Exp.	Class labels	EN	AR
Q1: Does the tweet contain a verifiable factual claim?		**504**	**218**
Bin	No	199	78
	Yes	305	140
Q2: To what extent does the tweet appear to contain false information?		**305**	**140**
Multi	No, definitely contains no false info	46	31
	No, probably contains no false info	177	62
	not sure	45	5
	Yes, probably contains false info	25	40
	Yes, definitely contains false info	12	2
Bin	No	223	93
	Yes	37	42
Q3: Will the tweet's claim have an effect on or be of interest to the general public?		**305**	**140**
Multi	No, definitely not of interest	10	1
	No, probably not of interest	46	5
	not sure	8	9
	Yes, probably of interest	180	76
	Yes, definitely of interest	61	49
Bin	No	56	6
	Yes	241	125
Q4: To what extent does the tweet appear to be harmful to society, person(s), company(s) or product(s)?		**305**	**140**
Multi	No, definitely not harmful	111	68
	No, probably not harmful	67	21
	not, sure	2	3
	Yes, probably harmful	67	46
	Yes, definitely harmful	58	2
Bin	No	178	89
	Yes	125	48
Q5: Do you think that a professional fact-checker should verify the claim in the tweet?		**305**	**140**
Multi	No, no need to check	81	22
	No, too trivial to check	64	55
	Yes, not urgent	117	48
	Yes, very urgent	43	15
Bin	No	145	77
	Yes	160	63
Q6: Is the tweet harmful for society and why?		**504**	**218**
Multi	No, joke or sarcasm	62	2
	No, not harmful	333	159
	not sure	2	0
	Yes, bad cure	3	1
	Yes, other	25	5
	Yes, panic	23	12
	Yes, rumor conspiracy	42	33
	Yes, xenophobic racist prejudices or hate speech	14	6
Bin	No	395	57
	Yes	107	218
Q7: Do you think that this tweet should get the attention of any government entity?		**504**	**218**
Multi	No, not interesting	319	163
	not sure	6	0
	Yes, asks question	2	0
	Yes, blame authorities	81	13
	Yes, calls for action	8	1
	Yes, classified as in question 6	34	30
	Yes, contains advice	9	1
	Yes, discusses action taken	12	6
	Yes, discusses cure	5	4
	Yes, other	28	0
Bin	No	319	163
	Yes	179	55

Table 2: Class distribution for both datasets. EN - English, AR - Arabic. In rows with question the number refers to the total number of tweets for the respective language. Bin - binary, Multi - multiclass. For the binary task, we re-label all *Yes** labels to **Yes** and *No** to **No**, and drop *not sure* labels.

the Arabic ones is similar, and therefore similar conclusions can be drawn.

We found that the class distribution for Q1 for the English tweets is quite balanced, (YES:61% and NO:39%). Only the tweets that are labeled as factual claims were annotated for Q2-5. For the question Q2, the label "NO, probably contains no false info" shows a higher distribution comparatively, which entails that in the majority of cases the identified claims are deemed to be most likely true. Out of 305 tweets labeled for Q2, in about 73% of the cases, it contains no false information, whereas 12% were categorized as "not sure" and 15% as "contains false information". While computing the statistics, we combined "probably" and "definitely" into one set for both positive and negative answers, respectively.

For Q3, which asks *if the tweet is of interest to general public*, we found the distribution to be skewed towards *Yes* with 79% of the distribution. This can be attributed to the fact that the tweets were selected based on frequency of retweets and likes. For Q4, which *judges if the tweet is harmful to the society*, the claims in the tweets vary from not harmful to harmful. For Q5 that asks *if a professional fact-checkers should verify the claim*, the majority of the cases were either "YES, not urgent" (38%) or "NO, no need to check" (27%). It appears that a professional fact-checker should verify the claims mentioned in the tweets immediately in only a small number of cases (14%). For Questions Q3-5, the "not sure" cases are generally very few. However, "not sure" cases are substantially more prevalent in the case of Q2. False information identification (Q2) is a challenging, because it requires further probing into external information. When annotating Q2, annotators were asked to examine the content of the tweets (i.e., user identifier, threads, videos, and images), by examining the whole tweet from its original URL. For example, in the following tweet *Epidemiologist Marc Lipsitch, director of Harvard's Center for Communicable Disease Dynamics: "In the US it is the opposite of contained.' https://t.co/IPAPagz4Vs"* it was difficult to determine whether it contains false information without looking at the tweet in its entirety. The original tweets often contain images and videos, which can help to identify the veracity of the claim and it is necessary to look at them.

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

4 Evaluation

We experiment with both binary and multiclass settings for our English and Arabic datasets (see Table 2 for data statistics).

4.1 Experimental Setup

Data Preprocessing The preprocessing includes removal of hash-symbol, non-ASCII characters; case folding; URLs are replaced with a URL tag; and usernames are also replaced with a user tag.

We split the data into 10-folds while maintaining the class distribution as close as possible to the overall distribution. For all tasks, we perform 10-fold cross-validation where for each run, we use train, dev, and test sets with with 80%, 10%, and 10% splits respectively.

Models Pre-trained models have achieved state-of-the-art performance for several NLP tasks. We experiment with several pre-trained models to evaluate their efficacy under various training scenarios such as, binary versus multiclass classification, low-resource task scenarios, presence of multilingual dataset, etc. More specifically, we use BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) for English language experiments and multilingual BERT (mBERT), XLM-r (Conneau et al., 2019) and AraBERT (Baly et al., 2020) for Arabic language experiments. In addition to pre-trained models, we also evaluate the performance of static-embedding based classification using FastText (Joulin et al., 2017).

For transformer-based models, we use the Transformer Toolkit (Wolf et al., 2019). We fine-tune each model using the default settings for three epochs as described in (Devlin et al., 2018). Due to the instability of the pre-trained models as reported by Devlin et al. (2018), we do 10 runs of each experiment using different random seeds and pick the model that performs the best on the development set. For FastText, we use pre-trained embeddings trained on Common Crawl (released by FastText for both English and Arabic). We also use the built-in hyperparameter tuning from the FastText

library to get the best set of hyperparameters on our development set.

Evaluation Metrics We compute the accuracy, precision (P), recall (R), and F1-measure (F1) (in terms of macro, and weighted average) from the overall confusion matrix of the 10-fold cross validation runs. Such a confusion matrix has been computed by summing up the individual confusion matrices. The reason to choose the weighted metric is that it takes into account the class imbalance problem. Due to the limited space, we report the weighted F1 scores in the paper. The results of other metrics are provided in the appendix H.4.

4.2 Results

4.2.1 Baseline

As a baseline, we use a majority class baseline i.e., class with the highest frequency. For questions with highly imbalanced class distributions, the majority class baseline is very high. For example, for Q3, we have an imbalanced distribution in the Arabic dataset, with tweets with a 'Yes' category in the binary setting comprised 93% of all tweets.

4.2.2 Binary Classification

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

Results on English dataset All models performed better than the majority class and FastText, confirming the efficacy of using contextualized embeddings in performing our proposed classification tasks. Comparing various pre-trained models, BERT outperformed all others in six out of seven tasks while ALBERT performed the worst in most of the cases. For Q1, RoBERTa and mBERT performed better than BERT with RoBERTa performing the best.

Results on Arabic dataset In all of the cases except Q3 Arabic, all models performed better than the majority class. For Q3, we have highly imbalanced distribution in the Arabic dataset with 125 instances of label 'Yes' and only 6 instances of label 'No'. Such imbalance complicates the task of machine learning models.

Comparing models, we did not see a clear domination of any of them as in the case of English. However, XLM-r had worse performance. mBERT outperformed all the other models in four out of seven tasks. AraBERT performed better than other models for Q4. Interestingly FastText achieved respectable performance on many tasks, with the best results on Q5. We hypothesize that due to the small size of Arabic data, the pre-trained models are likely to overfit and they may be less effective. Further, most pre-trained models are not trained on Arabic tweets.

4.2.3 Multiclass Classification

The second part of Table 3 presents the multiclass results. The *Cls* column shows the number of classes per task. Note that the size (i.e., the proportion of class labels) of the multiclass data is substantially smaller than the binary class. In addition, the multiclass is a much harder task compared to binary classification. This is reflected in the substantially lower classification results.

Results on English dataset All models performed better than the majority class. The most successful model is mBERT, which performed the best in four out of six tasks. It is interesting to see that mBERT outperformed BERT in several cases also. The size of data or the complexity of the task may have led to this result.

Results on Arabic dataset Interestingly, FastText outperformed all pre-trained models. None of the pre-trained models show consistent results on all tasks. This could be due to the scarcity of data, which may not be sufficient to optimize the large number of parameters of the pre-trained models. This is a useful finding that shows the value of static-embeddings in sparse data scenarios.

There are several possible interesting directions to explore in order to make pre-trained models effective in low-resource scenarios. For example, to avoid over-fitting instead of fine-tuning the whole model, one may fine-tune only a part of the network or fine-tune only the classification layer. Smaller BERT based models such as DistilBERT has significantly less number of parameters with a small drop in performance. Such might be effective in our case. We plan to explore this in future work.

4.2.4 Multilingual

Currently, the labeled datasets are comparatively small for both Arabic and English. In order to alleviate sparsity, we experiment with combining both Arabic and English data for training using a multilingual model. For training the models, we used mBERT and fine-tune the network for each question. The results are reported in Table 4.

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

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

The reported results suggest that overall we obtain better performance with multilingual training. For the binary task, Q2 monolingual results are higher, while multiclass results for Q2 and Q3 are higher.

Since simply concatenating the data of two languages is helpful, this encourages us to try other methods of data augmentation such as automatically translating the data of one language and using it with the target language. We left such exploration for future work.

Q.	Cls	English		Arabic	
		En	En+Ar	Ar	Ar+En
Binary (Coarse-grained)					
Q1	2	88.3	**89.7**	88.1	**89.9**
Q2	2	**83.1**	81.9	79.1	**82.5**
Q3	2	81.6	**83.0**	**89.2**	83.3
Q4	2	82.7	**83.2**	78.5	**86.0**
Q5	2	80.0	**83.9**	76.4	**78.6**
Q6	2	76.8	**83.0**	80.4	**82.8**
Q7	2	81.9	**84.0**	78.5	**84.6**
Multiclass (Fine-grained)					
Q2	5	**52.2**	47.0	42.8	**46.6**
Q3	5	45.1	**50.1**	27.0	**28.0**
Q4	5	42.9	**43.8**	43.7	**47.9**
Q5	4	52.3	**59.8**	59.0	**70.5**
Q6	8*	**62.7**	61.8	40.9	**42.7**
Q7	10*	**58.7**	58.4	66.3	**68.0**

Table 4: **Multilingual experiments using mBERT.** Shown are results for using Arabic data to help English, and English data to help Arabic (weighted F1).

5 Conclusion and Future Work

We presented an annotation scheme and a corresponding manually annotated dataset of COVID-19 tweets, aiming to help in the fight against the global infodemic, which emerged as a result of the COVID-19 pandemic. The dataset combines the perspectives and the interests of journalists, fact-checkers, social media platforms, policymakers, and society as a whole. It includes annotations in English and Arabic and is made freely available to the research community. We provided evaluation results for both English and Arabic using different Transformer model architectures.

We will be expanding the annotations, and we will make them available at (ANONYMOUS). We plan to recruit professional annotators to expand the size of the dataset significantly. We would also allow people to contribute to these annotations using a crowd-sourcing platform (ANONYMOUS).

There are a number of other interesting research directions that can be pursued using our dataset, such as multi-task learning (e.g., Q2, Q3, and Q4 can be used to improve the performance of Q5 as they are correlated; see Appendix D for more details), use of meta-data from Twitter (e.g., verified user account, reference to authentic URL, link to articles), of multi-modal information (e.g., image or video to verify the authenticity of the claim, retweets) for better classification and for data augmentation. Another research direction is to model the problem as ordinal regression, as the labels for questions Q2, Q3 and Q4 are defined on an ordinal scale.

References

Gerald Albaum. 1997. The likert scale revisited. *Market Research Society. Journal.*, 39(2):1–21.

Gordon W Allport and Leo Postman. 1947. The psychology of rumor.

Samah M Alzanin and Aqil M Azmi. 2018. Detecting rumors in social media: A survey. *Procedia computer science*, 142:294–300.

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 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Fady Baly, Hazem Hajj, et al. 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*, pages 9–15.

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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 21–27, New Orleans, LA, USA.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! 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.

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.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmn, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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.

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

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

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. 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.

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*, pages 845–854, Minneapolis, Minnesota, USA.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1835–1838.

Baani Leen Kaur Jolly, Palash Aggrawal, Amogh Gulati, Amarjit Singh Sethi, Ponnurangam Kumaraguru, and Tavpritesh Sethi. 2020. Psychometric analysis and coupling of emotions between state bulletins and twitter in india during covid-19 infodemic.

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: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

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

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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

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*.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16.

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.

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

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, MN, USA.

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

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 *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 372–387, Avignon, France. Springer.

David Pastor-Escuredo and Carlota Tarazona. 2020. Characterizing information leaders in twitter during covid-19 crisis.

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, pages 2259–2262, Singapore.

Andrew Perrin. 2015. Social media usage. *Pew research center*, pages 52–68.

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.

Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2020. An exploratory study of covid-19 misinformation on twitter.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

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 and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 3346–3359, Santa Fe, NM, USA.

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.

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)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

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

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. Detecting east asian prejudice on social media.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

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.

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

Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. 2020. Prevalence of low-credibility information on twitter during the covid-19 outbreak.

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. 2020. A first instagram dataset on covid-19.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Appendix

A Data Collection

The following keywords are used for collecting English and Arabic tweets.

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

B Detail Annotation Instructions

General Instructions:

1. For each tweet, the annotator needs to read the text including the hashtags and also 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).[4]
3. The annotators may look at the images and the videos, to the Web pages that the tweet links to, as well as to the tweets in the same thread when making a judgment, if required.
4. The annotators are not required to complete questions Q2-Q5 if the answer to question Q1 is **NO**.

B.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:[5]

- Stating a definition;
- Mentioning quantity in the present or the past;
- Making a verifiable prediction about the future;
- 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 which 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, 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 exist in a sub-sentence or sub-clause, then the tweet is not considered to have 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. Moreover, 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:

[4] 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 if 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. The annotator should consider the time when the tweet was posted.

[5] Inspired by (Konstantinovskiy et al., 2018).

- **`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 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. *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 its a must read for all those idiots clearing the shelves #coronavirus #toiletpapercrisis #auspol*
   Label: `NO`
   Explanation: There is no claim in the text.

B.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 the readers (Kumar and Shah, 2018).

Labels: The labels for this question are defined with 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 for the veracity of the claim identified in question 1. 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".

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/flSwhLCPmx*
   Label: `2. NO, probably contains no false info`
   Explanation: The content of the tweet states correct information.

B.3 Interest to General Public

Question 3: Will the tweet's claim have an effect on or be of interest to the general public?

Most often people do not make interesting claims, which can be verified by our general knowledge. For example, 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:

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.*

Its death rate is tiny bcos its 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 providing statement of someone else.

B.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 if 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*[6] (Broniatowski et al., 2018). A rumor involves a form of a statement whose veracity is not quickly verifiable or ever confirmed[7].

Labels: To categorize the tweets we defined the following labels based on the Likert scale. A higher value means 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`**

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 youre 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.

B.5 Need of Verification

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

It is important to verify a factual claim by a professional fact-checker, 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 worthwhile to be fact-checked by a professional fact-checker, because it is a time-consuming procedure. 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.

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, it is not urgent or critical;
4. **`YES, very urgent`**: the tweet can cause immediate harm to a large number of people,

[6] The use of information as a weapon to spread misinformation and mislead people.

[7] https://en.wikipedia.org/wiki/Rumor

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.

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 havent done: - Help workers - Help small businesses - Produced enough tests or ventilators*
   Label: `2. YES, very urgent`
   Explanation: Clearly, the content of the tweet blames authority, hence, it is important to verify this claim immediately 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 a joke and does not need to be checked by a professional fact checker.

B.6 Harmful to Society

Question 6: Is the tweet harmful for society and why?

The purpose of this question is to categorize if 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, itll 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 ov COVID-19 in Qatar"*).

E. **`YES, xenophobic, racist, prejudices, or hate-speech:`** the tweet reports xenophobia, racism or prejudiced expression(s). According to the dictionary[8] *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.[9] 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 of a social group.[10] An example of a xenophobic statement is *"do not buy cucumbers from Iran"*.

F. **`YES, bad cure:`** the tweet reports a questionable cure, medicine, vaccine or prevention procedures (e.g., *"... drinking bleach can help cure coronavirus"*).

G. **`YES, rumor, or conspiracy:`** the tweet reports or spreads a rumor. It 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.

[8] https://www.dictionary.com/
[9] https://en.wikipedia.org/wiki/Racism
[10] https://www.simplypsychology.org/prejudice.html

B.7 Requires attention

Question 7: Do you think that this tweet should get the attention of any government entity?

Most often people tweet by blaming authorities, providing advice, and/or call for action. Sometimes that information might be useful for some government entities to make a plan, respond or react on it. 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 that requires government attention. For the annotation, it is mandatory to first decide on whether attention is necessary from government entities (i.e., **`YES/NO`**). If the answer is **`YES`**, it is obligatory to select a category from the **`YES`** sub-categories mentioned below.

Labels:

A. **`NO, not interesting:`** if 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:`** if some government entities need to pay attention to this tweet as it is harmful for society, i.e., it is 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, blame authorities:`** the tweet contains information that blames some government entities or top politician(s), 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 entities (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 contains information that states that some government entities 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 gardgets keep in their houses. Salons and Barbershops might prove to be another Virus spreading channels @citizentvkenya @SenMutula @CSMutahi_Kagwe*).

H. **`YES, discusses action taken:`** if 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:`** if attention is needed from some government entities as the tweet discusses a possible cure, vaccine, or treatment for a disease (e.g., *Pls share this valuable information. Garlic boiled water can be cure corona virus*).

J. **`YES, asks question:`** if the content of the tweet contains a question over a particular issue and it requires attention from government entities (e.g., *Special thanks to all doctors and nurses, new found respect for youll. 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?*)

C Class Label Distribution

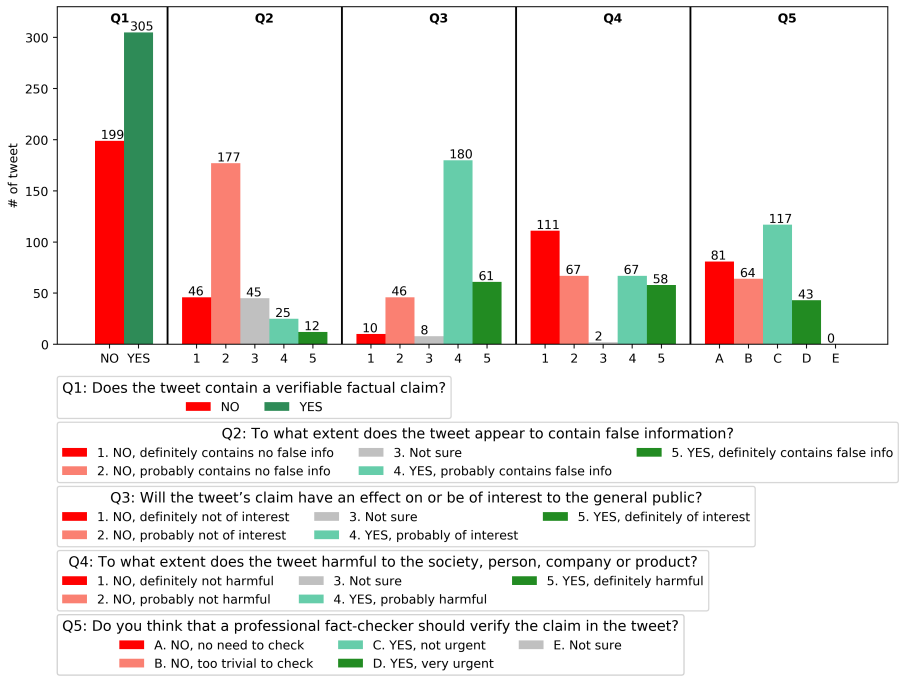
In Figure 2 and 3, we report detailed class label distribution of each question. In general the class distributions are similar in both English and Arabic.

D Correlation Between Questions

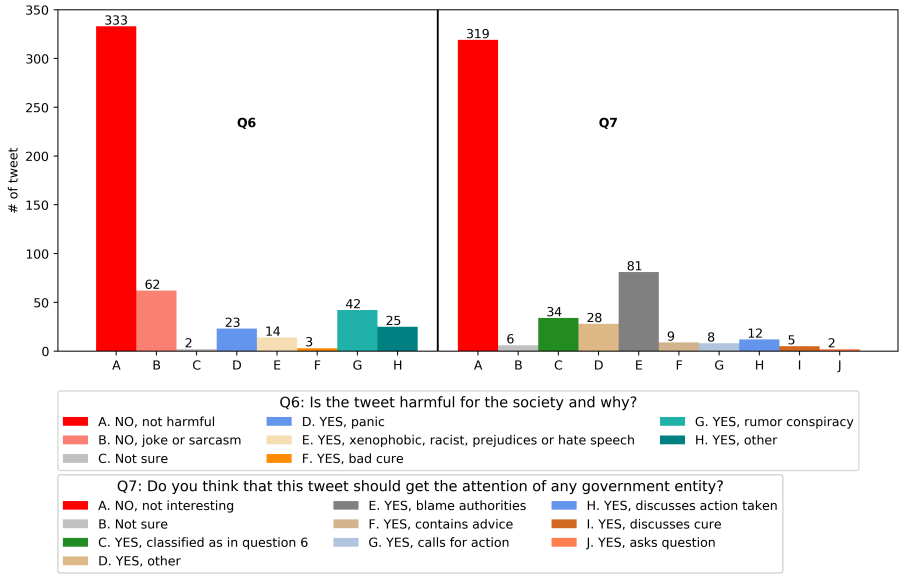
D.1 English Tweets Dataset

In Figure 4, we report the 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[11] between "...no false info"

[11]Note that, a Chi-Square test could have been a viable solution to prove such an association, however, our data size is still small (in many cases cell values are less than 5) to do such a test.

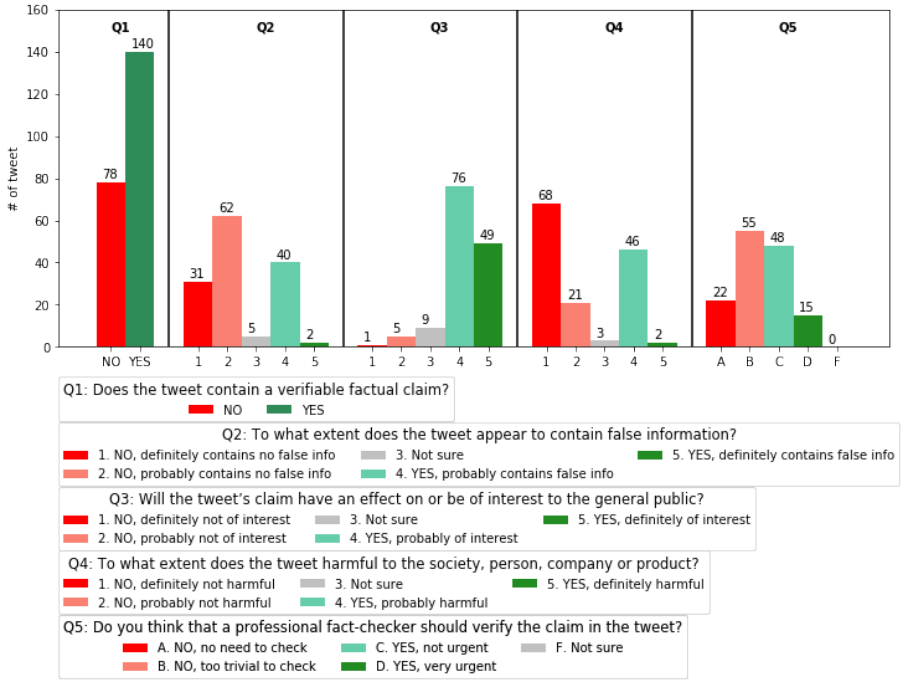


(a) Questions (Q1-5).

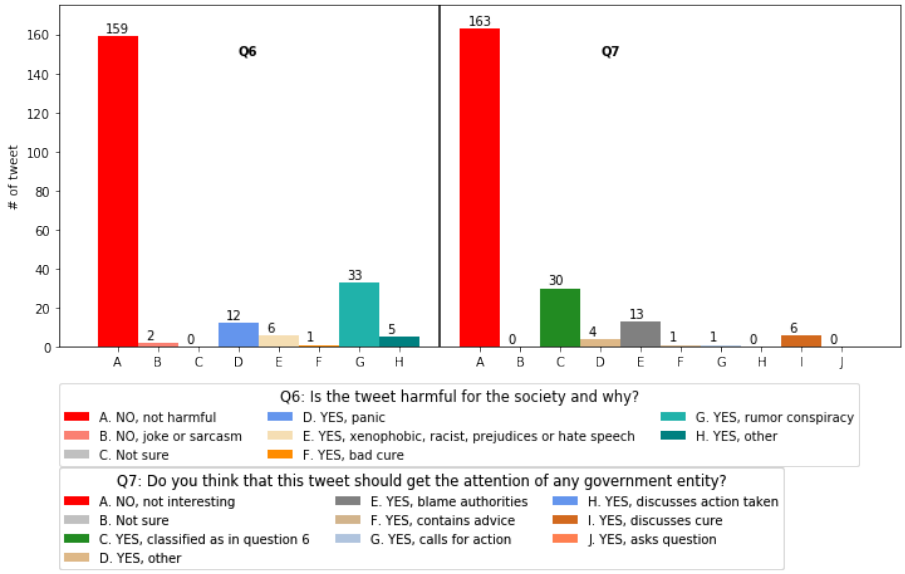


(b) Questions (Q6-7).

Figure 2: Distribution of class labels for **English tweets**

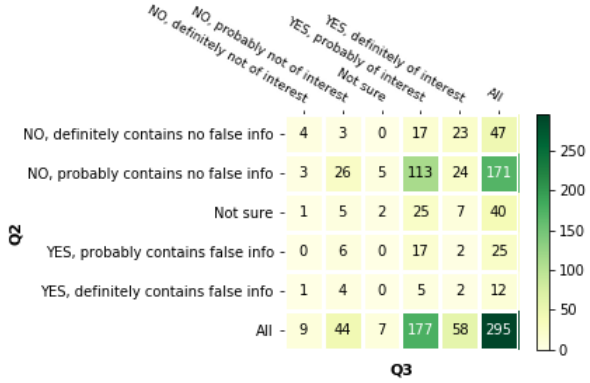


(a) Questions (Q1-5).

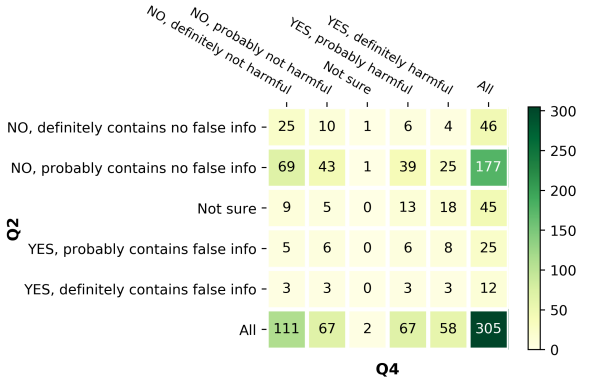


(b) Questions (Q6-7).

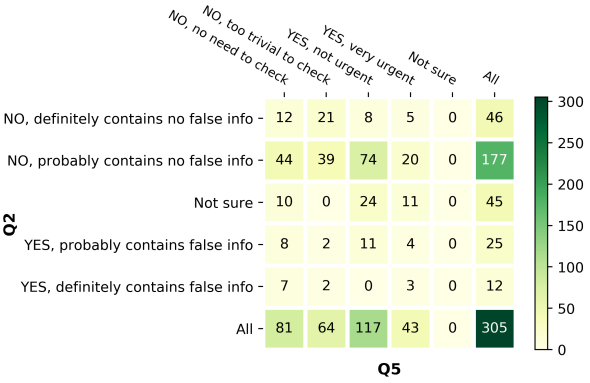
Figure 3: Distribution of class labels for **Arabic tweets**



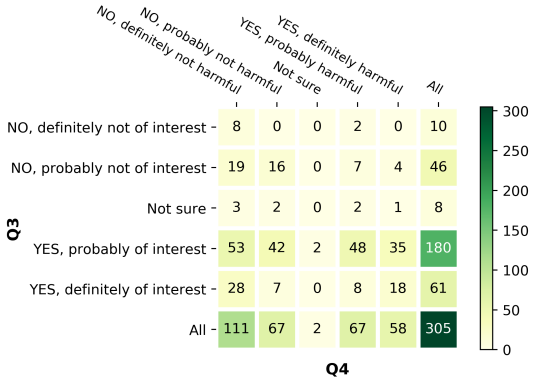
(a) Heatmap for Q2 and Q3.



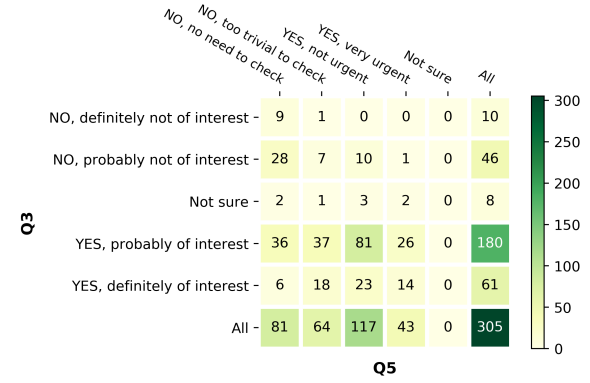
(b) Heatmap for Q2 and Q4.



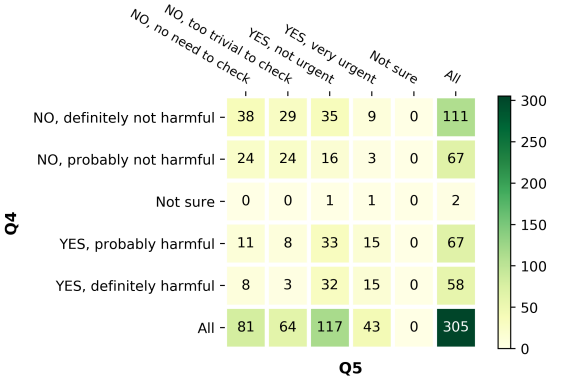
(c) Heatmap for Q2 and Q5.



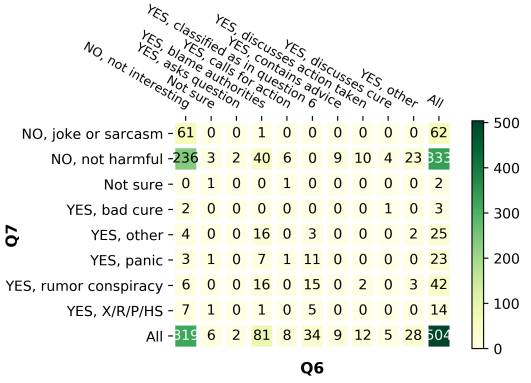
(d) Heatmap for Q3 and Q4.



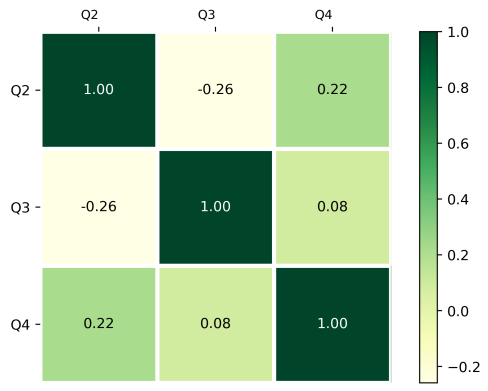
(e) Heatmap for Q3 and Q5.



(f) Heatmap for Q4 and Q5.

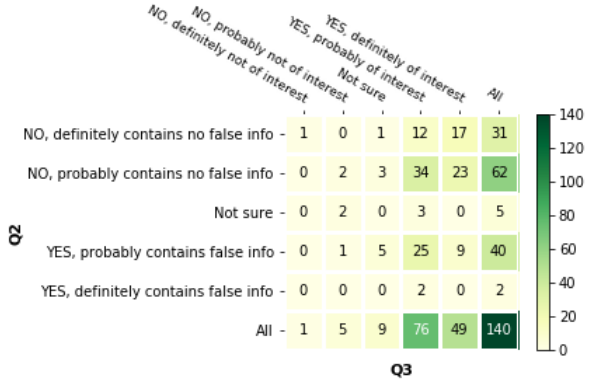


(g) Heatmap for Q6 and Q7. YES, X/R/P/HS – YES, xenophobic, racist, prejudices or hate speech

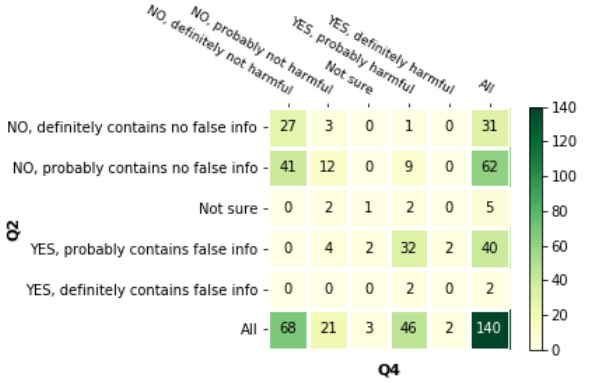


(h) Correlation between Q2 to Q4.

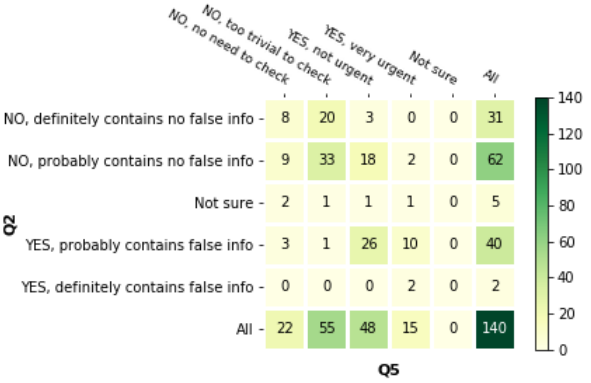
Figure 4: Contingency and correlation heatmaps of **English tweets** for different question pairs



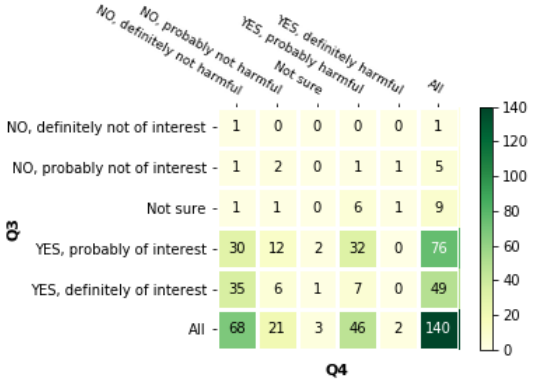
(a) Heatmap for Q2 and Q3.



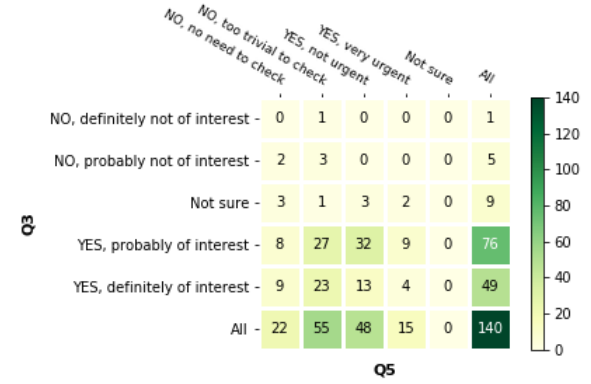
(b) Heatmap for Q2 and Q4.



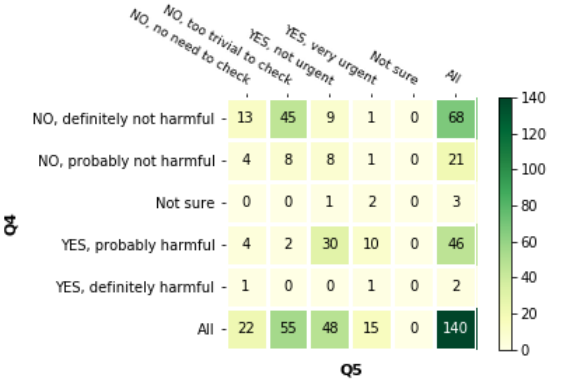
(c) Heatmap for Q2 and Q5.



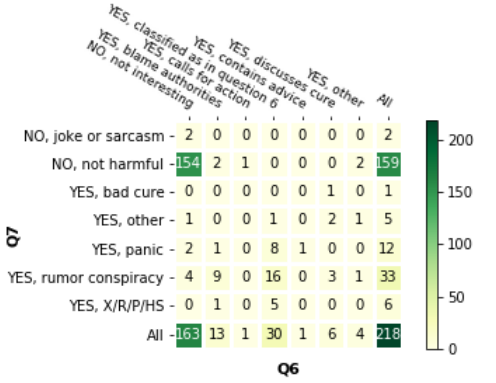
(d) Heatmap for Q3 and Q4.



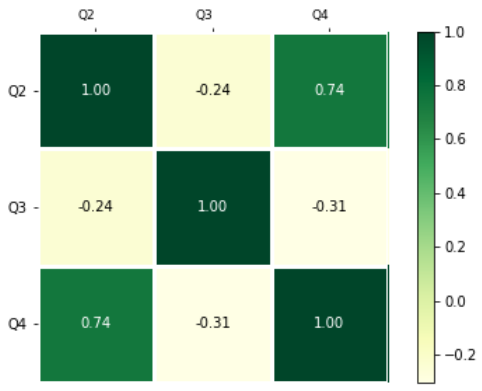
(e) Heatmap for Q3 and Q5.



(f) Heatmap for Q4 and Q5.



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



(h) Correlation between Q2 to Q4.

Figure 5: Contingency and correlation heatmaps of **Arabic tweets** for different question pairs

and the general public interest as shown in Figure 4a. For questions Q2 and Q4 (Figure 4b), a high association can be observed between "... no false info" and "... not harmful" (65%) compared to "harmful" (34%) for either an individual, products or government entities. By analyzing questions Q2 and Q5 (Figure 4c), 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 4d), it appears that when the content of the tweets is "not harmful" the general public interest is higher (61%) than when it is "harmful" (39%). From question Q3 and Q5 (Figure 4e), we see an interesting phenomenon, namely tweets with a high general public interest have a greater association with a professional fact-checker having to verify them (61%) compared to either "too trivial to check" or "no need to check" (39%). The questions Q4 and Q5 (Figure 4f) show that "harmful" tweets require more attention (53%) from a professional fact-checkers than "not harmful" tweets (45%). Our findings for Q6 and Q7 (Figure 4g) suggest that the majority of the tweets are not harmful for society, which also requires less attention from government entities. The second most common tweet label for Q7 blames authorities, though they are mostly not harmful for society.

In Figure 4h, we report the correlation between questions Q2-4 for the English tweets in order to understand their association. 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 infers that if the claim contains no false information, it is of high interest to the general public. This can be also observed in Figure 4a. Questions Q2 and Q4 show a positive correlation, which might be due to their high association with "... no false info" and "... not harmful".

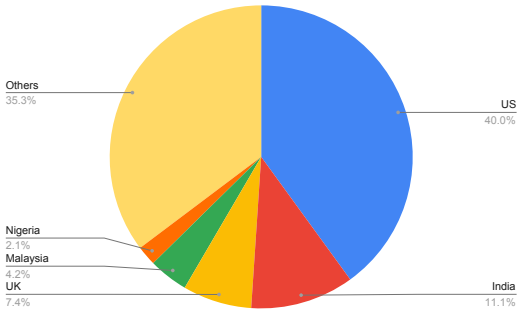
D.2 Arabic Tweets Dataset

In Figure 5, we report heatmaps to illustrate the association across questions using the Arabic tweets. From Q2 and Q3 (Figure 5a), we can observe that the association between "... contains no false info" and general public interest is higher (67%) than "... contains false info" (29%). From questions Q2 and Q4 (Figure 5b), we conclude that "... contains no false info" is associated with "... not harmful" and "... contains false info" is associated with "... harmful", which can also be established from its high correlation of 0.74 in Figure 5h. From the relation between Q2 and Q5 (Figure 5c), it can be seen 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 general public interest is higher when the content of the tweets is not harmful (68%) than harmful (30%) (Figure 5d). 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 5e). 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 5f). From the 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 5g).

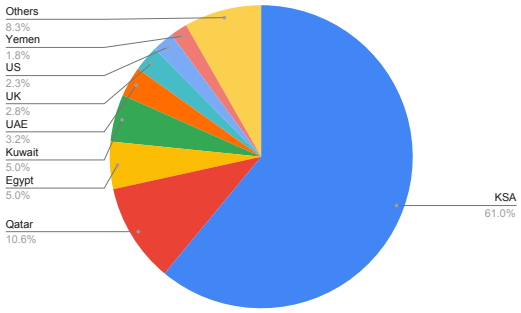
E Multimedia in Tweets

In this subsection, we study the correlation between whether a tweet has multimedia (video, image, or none) and our annotation. Generally, people trust videos more than images or plain texts which suggests that tweets with video potentially have a higher impact.

In the Arabic dataset, we didn't find any clear preference for Q1-Q4, i.e., a tweet with video, image, or only text can contain a factual claim with almost similar ratios. For Q5 (i.e., need fact-checking by a professional fact-checker), when a tweet has a video, in 44% of the cases, annotators selected "Yes" compared to 17% and 25% for image and text only respectively. Fact-checking of videos is not always trivial. For Q6 (i.e., harmful to society) and Q7 (i.e., should attract government attention), when a tweet has a video, it has the potential to be harmful and get government attention is higher than tweets with only images or text. Annotators selected "Yes" in almost 33% of the tweets having videos for Q6 and Q7 and this ratio decreased to almost half for tweets having images or text only. This shows the importance of having videos in tweets as it gives more trust.



(a) English dataset



(b) Arabic dataset

Figure 6: Country distribution for English and Arabic tweets

F Geographical distribution

Figure 6 shows the geographical distribution of annotated tweets for English and Arabic. We consider the country of the tweet author or the original author in case of retweeting. It is observed that most English tweets came from the US, India, and the UK (∼60%), while most Arabic tweets came from KSA and Qatar (∼70%). For both languages, there are tweets from a large number of countries, which indicates a good diversity of interests, topics, styles, etc. that strengthens our study.

G Verified and Unverified Accounts

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

Figure 7 shows that verified accounts tend to post more tweets that contain factual claims than unverified accounts (Q1), and their tweets are more likely to not contain false information (Q2), be of higher interest to the general public (Q3), be less harmful to society (Q6, Arabic), and attract greater attention from a government entity than tweets from unverified accounts (Q7, English). These are general observations from the currently small number of annotated tweets, and there are some differences between the English and Arabic annotations. The quantitative study can be held at a later stage using a larger dataset.

This correlation could be one of the features that a classifier can use to predict labels for unseen tweets, can also help in speeding up the annotation process by providing initial default values before manual revision. In addition, in some cases, verified accounts can be used to check annotation quality, for example, tweets from @WHO should not be labeled as weaponized or harmful to society.

H Experimental Parameters and Results

H.1 Transformers Parameters

Below we list the hyperparameters that we used for training across all Transformers based models. All experimental scripts will be publicly available.

- Batch size: 8
- Learning rate (Adam): 2e-5
- Number of epochs: 3
- Max seq length: 128

Number of parameters:

- **BERT** (bert-base-uncased): L=12, H=768, A=12, total parameters = 110M; where L is 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 with a higher number of parameters (125M).
- **ALBERT** (albert-base-v1): similar to BERT-base with a reduced parameters size of 12M.
- **AraBERT** (bert-base-arabert): same number as BERT (110M).
- **BERT Multilingual** (bert-base-multilingual-uncased) (mBERT): similar to BERT-base with a higher number of parameters (172M).
- **XML-RoBERTa** (xlm-roberta-base): L=12, H=768, A=12; total parameters = 270M.

H.2 FastText Parameters

We plan to release all the FastText parameters with our released packages. We have not listed them here due to their exhaustive list.

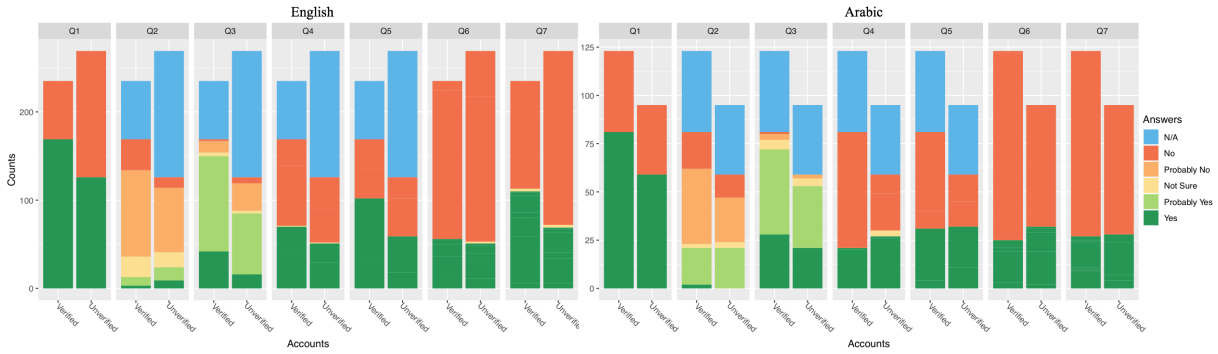


Figure 7: Distribution of datasets for all the questions associated with user accounts. NA refers to tweets that have not been labeled for those questions, they are identical to the tweets categorized with the label NO in Q1.

H.3 Computing Infrastructure and Runtime

We used the NVIDIA Tesla V100-SXM2-32 GB GPU machine consists of 56 cores and 256GB CPU memory. To perform an experiment for a question on average the computing time took 40 minutes using a BERT base model, which results in around 4 hours for seven questions using one transformer architecture.

H.4 Results

The detail classification results on dev and test sets in terms of accuracy (Acc), macro-F1 (M-F1) and weighted-F1 (W-F1) for English data are reported in Table 5 and 6, respectively. For Arabic data the detail of dev and test sets results are reported in Table 7 and 8, respectively.

In Table 9 and 10, we report on dev and test set for multilingual setting where training is performed by combining English and Arabic data and evaluated on English data. With the same multilingual setting the results on Arabic evaluation is reported in Table 11 and 12 for dev and test sets, respectively.

	Binary			Multiclass		
Q.	**Acc.**	**M-F1**	**W-F1**	**Acc.**	**M-F1**	**W-F1**
			Majority			
Q1	60.8	37.8	46.0	-	-	-
Q2	85.2	46.0	78.4	51.1	18.9	44.0
Q3	80.0	44.4	71.1	54.1	22.5	48.3
Q4	58.1	36.7	42.7	36.4	27.5	35.5
Q5	51.6	34.0	35.1	38.7	33.2	37.6
Q6	78.4	44.0	69.0	63.9	12.60	53.9
Q7	64.0	39.0	50.0	64.9	14.00	57.8
			BERT			
Q1	92.0	91.5	91.9	-	-	-
Q2	88.1	66.0	85.3	63.5	24.3	52.5
Q3	90.3	81.2	89.1	58.1	28.1	54.1
Q4	90.0	89.7	90.0	51.9	33.4	44.5
Q5	89.0	89.0	89.0	61.9	51.8	58.0
Q6	92.5	88.0	92.2	69.4	15.5	59.2
Q7	93.4	92.8	93.4	62.9	10.70	57.5
			RoBERTa			
Q1	93.7	93.4	93.7	-	-	-
Q2	87.0	60.7	83.3	64.5	36.0	55.5
Q3	87.3	76.0	85.9	54.2	28.7	52.4
Q4	91.3	91.0	91.2	51.9	35.1	47.6
Q5	85.5	85.4	85.4	67.7	56.9	64.1
Q6	89.2	82.8	88.8	72.2	20.1	64.2
Q7	90.8	89.9	90.7	61.6	11.40	57.6
			ALBERT			
Q1	91.8	91.2	91.7	-	-	-
Q2	87.8	66.6	85.3	60.0	20.2	47.4
Q3	89.0	77.8	87.3	52.6	24.8	49.3
Q4	85.8	85.4	85.8	47.1	31.0	41.8
Q5	82.9	82.9	82.9	68.1	57.6	64.6
Q6	86.3	74.6	84.4	71.6	17.6	62.1
Q7	89.2	88.4	89.2	63.1	9.70	54.7
			FastText			
Q1	74.3	72.2	73.8	-	-	-
Q2	85.9	56.6	81.7	60.6	26.6	51.8
Q3	80.0	55.8	75.4	55.5	23.2	48.9
Q4	75.5	74.7	75.4	43.5	32.4	42.3
Q5	67.1	66.6	66.7	45.2	41.7	45.2
Q6	77.3	55.6	73.2	68.0	15.9	58.8
Q7	75.6	71.6	74.6	68.4	16.5	61.8

Table 5: Classification results on **dev set (English data)** using different models including majority baseline for different questions. Acc. - Accuracy, M-F1 - macro F1, W-F1 - weighted average F1. For Q1, binary and multiclass setting's results are same.

	Binary			Multiclass		
Q.	**Acc.**	**M-F1**	**W-F1**	**Acc.**	**M-F1**	**W-F1**
			Majority			
Q1	60.5	37.7	45.6	-	-	-
Q2	85.8	46.2	79.2	58.0	14.7	42.6
Q3	81.1	44.8	72.7	59.0	14.8	43.8
Q4	58.7	37.0	43.5	36.4	10.7	19.4
Q5	52.5	34.4	36.1	38.4	13.9	21.3
Q6	78.7	44.0	69.3	66.1	9.90	52.6
Q7	64.1	39.0	50.0	63.3	7.80	49.1
			BERT			
Q1	87.7	87.0	87.6	-	-	-
Q2	89.2	69.0	86.9	59.7	21.9	48.5
Q3	86.5	71.1	84.3	60.3	30.6	57.6
Q4	84.2	83.3	84.0	49.2	29.5	41.6
Q5	81.3	81.2	81.3	54.8	44.6	50.4
Q6	87.1	77.9	86.1	68.3	14.0	57.2
Q7	89.4	88.4	89.3	62.7	9.50	54.6
			RoBERTa			
Q1	90.7	90.1	90.6	-	-	-
Q2	86.9	57.8	82.9	58.4	21.3	46.6
Q3	83.5	65.0	80.8	52.5	25.9	50.9
Q4	83.8	83.2	83.8	49.2	31.8	44.1
Q5	73.8	73.6	73.7	54.4	43.2	50.3
Q6	82.3	70.0	81.0	67.5	15.6	58.4
Q7	84.9	83.2	84.7	59.3	9.70	55.2
			ALBERT			
Q1	86.5	85.8	86.5	-	-	-
Q2	87.7	60.3	83.9	58.7	17.3	44.8
Q3	83.8	61.1	79.6	50.5	19.6	45.4
Q4	78.5	77.7	78.5	45.9	31.0	39.5
Q5	72.8	72.6	72.7	52.1	41.4	48.0
Q6	82.5	65.1	79.2	66.9	13.9	56.5
Q7	79.3	76.8	79.0	61.5	9.80	53.5
			FastText			
Q1	73.2	71.1	72.8	-	-	-
Q2	86.5	57.4	82.6	51.1	18.9	44.0
Q3	80.5	58.1	77.2	54.1	22.5	48.3
Q4	70.3	68.1	69.6	36.4	27.5	35.5
Q5	63.3	62.9	63.1	38.7	33.2	37.6
Q6	75.7	52.7	71.6	63.9	12.6	53.9
Q7	70.5	66.8	69.9	64.9	14.0	57.8

Table 6: Classification results on **test set (English data)** using different models including majority baseline for different questions.

	Binary			Multiclass		
Q.	**Acc.**	**M-F1**	**W-F1**	**Acc.**	**M-F1**	**W-F1**
			Majority			
Q1	63.6	38.9	49.5	-	-	-
Q2	71.4	41.7	59.5	44.3	12.3	27.2
Q3	92.9	48.1	89.4	54.3	14.1	38.2
Q4	64.3	39.1	50.3	48.6	13.1	31.8
Q5	53.3	34.8	37.1	39.3	14.1	22.2
Q6	72.7	42.1	61.2	72.9	12.1	61.5
Q7	75.0	42.9	64.3	74.8	12.2	64.0
			mBERT			
Q1	88.2	87.5	88.3	-	-	-
Q2	86.4	82.1	85.9	54.7	30.1	49.8
Q3	84.3	49.9	85.5	24.0	15.2	30.0
Q4	85.0	82.5	84.4	52.0	23.7	44.4
Q5	83.3	83.3	83.4	74.0	48.5	65.2
Q6	84.1	78.7	83.6	31.4	11.0	42.0
Q7	85.9	76.1	83.7	82.3	18.4	75.5
			AraBERT			
Q1	84.5	82.5	84.1	-	-	-
Q2	84.3	78.6	83.3	54.0	24.1	45.7
Q3	67.1	45.7	74.9	19.3	12.8	23.5
Q4	86.4	84.2	85.9	56.0	24.2	51.1
Q5	82.7	82.2	82.4	72.7	51.5	65.5
Q6	86.4	80.0	85.1	30.0	10.9	40.7
Q7	86.4	77.1	84.4	81.4	16.5	74.2
			XML-r			
Q1	79.1	74.9	77.7	-	-	-
Q2	75.7	57.3	69.3	47.3	17.0	34.8
Q3	84.3	49.9	85.5	22.0	14.4	25.5
Q4	76.4	67.6	72.4	46.0	15.5	34.1
Q5	68.7	65.6	66.3	63.3	36.1	53.3
Q6	74.1	48.6	65.0	37.7	8.8	43.9
Q7	75.0	42.9	64.3	80.5	12.7	71.7
			FastText			
Q1	73.6	69.1	72.4	-	-	-
Q2	85.0	81.2	84.8	63.3	38.1	61.2
Q3	92.9	48.1	89.4	84.7	58.8	84.0
Q4	82.1	78.3	80.9	68.0	34.2	64.5
Q5	80.0	79.9	80.0	77.3	66.9	75.2
Q6	82.3	74.6	80.9	77.3	26.0	72.4
Q7	85.5	76.5	83.7	81.4	20.4	76.1

Table 7: Classification results on **dev set (Arabic data)** using different models including majority baseline for different questions.

	Binary			Multiclass		
Q.	**Acc**	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**
			Majority			
Q1	64.2	39.1	50.2	-	-	-
Q2	68.9	40.8	56.2	44.3	12.3	27.2
Q3	95.4	48.8	93.2	54.3	14.1	38.2
Q4	65.0	39.4	51.2	48.6	13.1	31.8
Q5	55.0	35.5	39.0	39.3	14.1	22.2
Q6	73.9	42.5	62.7	72.9	12.1	61.5
Q7	74.8	42.8	64.0	74.8	12.2	64.0
			mBERT			
Q1	88.1	87.0	88.1	-	-	-
Q2	80.0	74.7	79.1	48.6	25.1	42.8
Q3	87.0	51.8	89.2	22.9	14.5	27.0
Q4	78.8	76.1	78.5	52.9	20.1	43.7
Q5	76.4	76.2	76.4	65.7	44.5	59.0
Q6	81.7	73.3	80.4	30.7	10.5	40.9
Q7	80.7	69.0	78.5	75.7	15.4	66.3
			AraBERT			
Q1	83.0	80.6	82.6	-	-	-
Q2	74.1	63.8	71.1	50.0	23.2	42.1
Q3	68.7	45.0	77.8	17.9	11.1	21.4
Q4	81.0	78.0	80.4	53.6	20.6	44.9
Q5	76.4	75.7	76.1	65.0	42.6	57.7
Q6	79.8	68.1	77.3	28.9	9.8	38.9
Q7	80.7	67.9	77.9	74.3	12.2	63.9
			XML-r			
Q1	78.9	73.6	76.9	-	-	-
Q2	69.6	47.2	60.2	48.6	19.3	37.4
Q3	87.0	51.8	89.2	20.0	14.3	20.0
Q4	73.0	63.5	69.0	45.7	14.7	34.2
Q5	68.6	65.5	66.5	55.0	31.2	46.1
Q6	74.3	45.9	64.6	38.5	9.0	44.5
Q7	74.8	42.8	64.0	74.8	12.2	64.0
			FastText			
Q1	76.6	73.1	75.8	-	-	-
Q2	69.6	61.5	68.2	49.3	29.2	47.4
Q3	95.4	48.8	93.2	83.6	64.5	83.1
Q4	79.6	76.8	79.2	57.1	28.9	54.4
Q5	78.6	78.5	78.6	78.6	70.6	77.2
Q6	81.2	71.5	79.4	82.6	35.5	79.3
Q7	77.1	62.5	74.1	80.3	33.6	75.7

Table 8: Classification results on **test set (Arabic data)** using different models including majority baseline for different questions.

Q.	**Acc.**	**M-F1**	**W-F1**	**Acc**	**M-F1**	**W-F1**
	Binary: En			**Binary: En+Ar**		
Q1	91.2	90.6	91.1	92.9	92.5	92.9
Q2	89.6	71.2	87.4	87.8	63.0	84.3
Q3	88.0	77.3	86.6	88.0	78.4	87.0
Q4	90.6	90.3	90.6	89.4	89.2	89.4
Q5	85.2	85.2	85.2	85.8	85.6	85.7
Q6	84.9	71.9	82.8	90.2	83.8	89.6
Q7	91.6	90.9	91.6	91.6	91.0	91.6
	Multiclass: En			**Multiclass: En+Ar**		
Q2	63.5	24.3	52.5	61.9	27.7	50.9
Q3	58.1	28.1	54.1	57.4	29.1	53.3
Q4	51.9	33.4	44.5	48.1	35.2	47.8
Q5	61.9	51.8	58.0	73.9	70.0	72.8
Q6	69.4	15.5	59.2	72.9	22.8	66.6
Q7	62.9	10.7	57.5	64.7	11.8	59.8

Table 9: **Multilingual experiments:** classification results on **dev set** using mBERT for both binary and multiclass settings. For En+Ar setting, the training is performed by combining English and Arabic data and **evaluated on English data**.

Q.	**Acc.**	**M-F1**	**W-F1**	**Acc.**	**M-F1**	**W-F1**
	Binary: En			**Binary: En+Ar**		
Q1	87.7	87.0	87.6	89.9	89.1	89.7
Q2	89.2	69.0	86.9	86.2	55.3	81.9
Q3	86.5	71.1	84.3	84.8	69.7	83.0
Q4	84.2	83.3	84.0	83.2	82.7	83.2
Q5	81.3	81.2	81.3	83.9	83.8	83.9
Q6	87.1	77.9	86.1	84.3	72.9	83.0
Q7	89.4	88.4	89.3	83.9	82.8	84.0
	Multiclass: En			**Multiclass: En+Ar**		
Q1	87.7	87.0	87.6	89.9	89.1	89.7
Q2	59.7	21.9	48.5	59.3	20.4	47.0
Q3	60.3	30.6	57.6	54.8	25.0	50.1
Q4	49.2	29.5	41.6	44.9	34.5	43.8
Q5	54.8	44.6	50.4	61.3	56.5	59.8
Q6	68.3	14.0	57.2	68.5	19.8	61.8
Q7	62.7	9.5	54.6	62.3	11.6	58.4

Table 10: **Multilingual experiments:** classification results on **English test set** using mBERT for both binary and multiclass settings.

Q.	**Acc.**	**M-F1**	**W-F1**	**Acc.**	**M-F1**	**W-F1**
	Binary: Ar			**Binary: En+Ar**		
Q1	88.2	87.5	88.3	86.8	85.6	86.7
Q2	86.4	82.1	85.9	88.6	86.9	88.9
Q3	84.3	49.9	85.5	75.7	48.2	80.6
Q4	85.0	82.5	84.4	90.0	88.5	89.7
Q5	83.3	83.3	83.4	82.0	81.9	82.0
Q6	84.1	78.7	83.6	91.4	88.9	91.3
Q7	85.9	76.1	83.7	88.6	83.5	88.1
	Multiclass: Ar			**Multiclass: En+Ar**		
Q2	54.7	30.1	49.8	58.7	32.2	55.5
Q3	24.0	15.2	30.0	24.7	12.3	25.7
Q4	52.0	23.7	44.4	50.7	29.1	53.3
Q5	74.0	48.5	65.2	78.7	65.1	75.1
Q6	31.4	11.0	42.0	32.3	10.3	43.3
Q7	82.3	18.4	75.5	81.4	12.8	77.1

Table 11: **Multilingual experiments:** classification results on **dev set** using mBERT for both binary and multiclass settings. For En+Ar setting, the training is performed by combining English and Arabic data and **evaluated on Arabic data**

Q.	Acc	M-F1	W-F1	Acc	M-F1	W-F1
	Binary: Ar			**Binary: En+Ar**		
Q1	88.1	87.0	88.1	89.9	89.0	89.9
Q2	80.0	74.7	79.1	82.2	80.0	82.5
Q3	87.0	51.8	89.2	77.1	46.6	83.3
Q4	78.8	76.1	78.5	86.1	84.5	86.0
Q5	76.4	76.2	76.4	78.6	78.4	78.6
Q6	81.7	73.3	80.4	83.0	77.4	82.8
Q7	80.7	69.0	78.5	85.3	78.6	84.6
	Multiclass: Ar			**Multiclass: En+Ar**		
Q2	48.6	25.1	42.8	51.4	26.5	46.6
Q3	22.9	14.5	27.0	27.9	14.8	28.0
Q4	52.9	20.1	43.7	51.4	23.0	47.9
Q5	65.7	44.5	59.0	73.6	59.7	70.5
Q6	30.7	10.5	40.9	32.1	10.1	42.7
Q7	75.7	15.4	66.3	73.9	10.9	68.0

Table 12: **Multilingual experiments:** classification results on **Arabic test set** using mBERT for both binary and multiclass settings.