Findings of the NLP4IF-2021 Shared Tasks on Fighting the COVID-19 Infodemic and Censorship Detection

Shaden Shaar,[1] Firoj Alam,[1] Giovanni Da San Martino,[2] Alex Nikolov,[3]
Wajdi Zaghouani,[4] Preslav Nakov, [1] and Anna Feldman[5]
[1] Qatar Computing Research Institute, HBKU, Qatar
[2] University of Padova, Italy, [3] Sofia University "St. Kliment Ohridski", Bulgaria,
[4] Hamad bin Khalifa University, Qatar, [5] Montclair State University, USA
{sshaar,fialam,wzaghouani,pnakov}@hbku.edu.qa
dasan@math.unipd.it feldmana@montclair.edu

Abstract

We present the results and the main findings of the NLP4IF-2021 shared tasks. Task 1 focused on fighting the COVID-19 infodemic in social media, and it was offered in Arabic, Bulgarian, and English. Given a tweet, it asked to predict whether that tweet contains a verifiable claim, and if so, whether it is likely to be false, is of general interest, is likely to be harmful, and is worthy of manual fact-checking; also, whether it is harmful to society, and whether it requires the attention of policy makers. Task 2 focused on censorship detection, and was offered in Chinese. A total of ten teams submitted systems for task 1, and one team participated in task 2; nine teams also submitted a system description paper. Here, we present the tasks, analyze the results, and discuss the system submissions and the methods they used. Most submissions achieved sizable improvements over several baselines, and the best systems used pre-trained Transformers and ensembles. The data, the scorers and the leaderboards for the tasks are available at http://gitlab.com/NLP4IF/nlp4if-2021.

1 Introduction

Social media have become a major communication channel, enabling fast dissemination and consumption of information. A lot of this information is true and shared in good intention; however, some is false and potentially harmful. While the so-called "fake news" is not a new phenomenon, e.g., the term was coined five years ago, the COVID-19 pandemic has given rise to the first global social media infodemic. The infodemic has elevated the problem to a whole new level, which goes beyond spreading fake news, rumors, and conspiracy theories, and extends to promoting fake cure, panic, racism, xenophobia, and mistrust in the authorities, among others. Identifying such false and potentially malicious information in tweets is important to journalists, fact-checkers, policy makers, government entities, social media platforms, and society.

A number of initiatives have been launched to fight this infodemic, e.g., by building and analyzing large collections of tweets, their content, source, propagators, and spread (Leng et al., 2021; Medford et al., 2020; Mourad et al., 2020; Karami et al., 2021). Yet, these efforts typically focus on a specific aspect, rather than studying the problem from a holistic perspective. Here we aim to bridge this gap by introducing a task that asks to predict whether a tweet contains a verifiable claim, and if so, whether it is likely to be false, is of general interest, is likely to be harmful, and is worthy of manual fact-checking; also, whether it is harmful to society, and whether it requires the attention of policy makers. The task follows an annotation schema proposed in (Alam et al., 2020, 2021b).

While the COVID-19 infodemic is characterized by insufficient attention paid to the problem, there are also examples of the opposite: tight control over information. In particular, freedom of expression in social media has been supercharged by a new and more effective form of digital authoritarianism. Political censorship exists in many countries, whose governments attempt to conceal or to manipulate information to make sure their citizens are unable to read or to express views that are contrary to those of people in power. One such example is Sina Weibo, a Chinese microblogging website with over 500 million monthly active users, which sets strict control over its content using a variety of strategies to target censorable posts, ranging from keyword list filtering to individual user monitoring: among all posts that are eventually censored, nearly 30% are removed within 5–30 minutes, and for 90% this is done within 24 hours (Zhu et al., 2013). We hypothesize that the former is done automatically, while the latter involves human censors. Thus, we propose a shared task that aims to study the potential for automatic sensorship, which asks participating systems to predict whether a Sina Weibo post will be censored.

2 Related Work

In this section, we discuss studies relevant to the COVID-19 infodemic and to censorship detection.

2.1 COVID-19 Infodemic

Disinformation, misinformation, and "fake news" thrive in social media. Lazer et al. (2018) and Vosoughi et al. (2018) in Science provided a gen-eral discussion on the science of "fake news" and the process of proliferation of true and false news online. There have also been several interesting surveys, e.g., Shu et al. (2017) studied how infor-mation is disseminated and consumed in social media. Another survey by Thorne and Vlachos (2018) took a fact-checking perspective on "fake news" and related problems. Yet another survey (Li et al., 2016) covered truth discovery in gen-eral. Some very recent surveys focused on stance for misinformation and disinformation detection (Hardalov et al., 2021), on automatic fact-checking to assist human fact-checkers (Nakov et al., 2021a), on predicting the factuality and the bias of entire news outlets (Nakov et al., 2021c), on multimodal disinformation detection (Alam et al., 2021a), and on abusive language in social media (Nakov et al., 2021b).

A number of Twitter datasets have been developed to address the COVID-19 infodemic. Some are without labels, other use distant supervision, and very few are manually annotated. Cinelli et al. (2020) studied COVID-19 rumor amplification in five social media platforms; their data was labeled using distant supervision. Other datasets include a multi-lingual dataset of 123M tweets (Chen et al., 2020), another one of 383M tweets (Banda et al., 2020), a billion-scale dataset of 65 languages and 32M geo-tagged tweets (Abdul-Mageed et al., 2021), and the GeoCoV19 dataset, consisting of 524M multilingual tweets, including 491M with GPS coordinates (Qazi et al., 2020). There are also Arabic datasets, both with (Haouari et al., 2021; Mubarak and Hassan, 2021) and without manual annotations (Alqurashi et al., 2020). We are not aware of Bulgarian datasets.

Zhou et al. (2020) created the ReCOVery dataset, which combines 2,000 news articles about COVID-19, annotated for their factuality, with 140,820 tweets. Vidgen et al. (2020) studied COVID-19 prejudices using a manually labeled dataset of 20K tweets with the following labels: hostile, criticism, prejudice, and neutral.

Song et al. (2021) collected a dataset of false and misleading claims about COVID-19 from IFCN Poynter, which they manually annotated with the following ten disinformation-related categories: (1) Public authority, (2) Community spread and impact, (3) Medical advice, self-treatments, and virus effects, (4) Prominent actors, (5) Conspiracies, (6) Virus transmission, (7) Virus origins and properties, (8) Public reaction, and (9) Vaccines, medical treatments, and tests, and (10) Cannot determine.

Another related dataset study by (Pulido et al., 2020) analyzed 1,000 tweets and categorized them based on factuality into the following categories: (i) False information, (ii) Science-based evidence, (iii) Fact-checking tweets, (iv) Mixed information, (v) Facts, (vi) Other, and (vii) Not valid. Ding et al. (2020) have a position paper discussing the challenges in combating the COVID-19 infodemic in terms of data, tools, and ethics. Hossain et al. (2020) developed the COVIDLies dataset by matching a known misconceptions with tweets, and manually annotated the tweets with stance: whether the target tweet agrees, disagrees, or has no position with respect to a known misconception. Finally, (Shuja et al., 2020) provided a comprehensive survey categorizing the COVID-19 literature into four groups: diagonisis related, transmission and mobility, social media analysis, and knowledge-based approaches.

The most relevant previous work is (Alam et al., 2021b, 2020), where tweets about COVID-19 in Arabic and English were annotated based on an annotation schema of seven questions. Here, we adopt the same schema (but with binary labels only), but we have a larger dataset for Arabic and English, and we further add an additional language: Bulgarian.

2.2 Censorship Detection

There has been a lot of research aiming at developing strategies to detect and to evade censorship. Most work has focused on exploiting technological limitations with existing routing protocols (Leberknight et al., 2012; Katti et al., 2005; Levin et al., 2015; Weinberg et al., 2012; Bock et al., 2020). Research that pays more attention to the linguistic properties of online censorship in the context of censorship evasion includes Safaka et al. (2016), who applied linguistic steganography to circumvent censorship.

Other related work is that of Lee (2016), who used parodic satire to bypass censorship in China and claimed that this stylistic device delays and often evades censorship. Hiruncharoenvate et al. (2015) showed that the use of homophones of censored keywords on Sina Weibo could help extend the time for which a Weibo post could remain available online. All these methods require significant human effort to interpret and to annotate texts to evaluate the likelihood of censorship, which might not be practical to carry out for common Internet users in real life.

King et al. (2013) in turn studied the relationship between political criticism and the chance of censorship. They came to the conclusion that posts that have a Collective Action Potential get deleted by the censors even if they support the state. Zhang and Pan (2019) introduced a system, Collective Action from Social Media (CASM), which uses convolutional neural networks on image data and recurrent neural networks with long short-term memory on text data in a two-stage classifier to identify social media posts about offline collective action. Zhang and Pan (2019) found that despite online censorship in China suppressing the discussion of collective action in social media, censorship does not have a large impact on the number of collective action posts identified through CASM-China. Zhang and Pan (2019) claimed that the system would miss collective action taking place in ethnic minority regions, such as Tibet and Xinjiang, where social media penetration is lower and more stringent Internet control is in place, e.g., Internet blackouts.

Finally, there has been research that uses linguistic and content clues to detect censorship. Knockel et al. (2015) and Zhu et al. (2013) proposed detection mechanisms to categorize censored content and to automatically learn keywords that get censored. Bamman et al. (2012) uncovered a set of politically sensitive keywords and found that the presence of some of them in a Weibo blogpost contributed to a higher chance of the post being censored. Ng et al. (2018b) also targeted a set of topics that had been suggested to be sensitive, but unlike Bamman et al. (2012), they covered areas not limited to politics. Ng et al. (2018b), Ng et al. (2019), and Ng et al. (2020) investigated how the textual content might be relevant to censorship decisions when both censored and uncensored blogposts include the same sensitive keyword(s).

3 Tasks

Below, we describe the two tasks: their setup and their corresponding datasets.

3.1 Task 1: COVID-19 Infodemic

Task Setup: The task asks to predict several binary properties for an input tweet about COVID-19. These properties are formulated in seven questions as briefly discussed below:

1. Verifiable Factual Claim: Does the tweet contain a verifiable factual claim? A verifiable factual claim is a statement that something is true, and this can be verified using factual, verifiable information such as statistics, specific examples, or personal testimony. Following (Konstantinovskiy et al., 2018), factual claims could be (a) stating a definition, (b) mentioning a quantity in the present or in the past, (c) making a verifiable prediction about the future, (d) reference laws, procedures, and rules of operation, and (e) reference images or videos (e.g., "This is a video showing a hospital in Spain."), (f) implying correlation or causation (such correlation/-causation needs to be explicit).

2. False Information: To what extent does the tweet appear to contain false information? This annotation determines how likely the tweet is to contain false information without fact-checking it, but looking at things like its style, metadata, and the credibility of the sources cited, etc.

3. Interesting for the General Public: Will the tweet have an impact on or be of interest to the general public? In general, claims about topics such as healthcare, political news and findings, and current events are of higher interest to the general public. Not all claims should be fact-checked, for example "The sky is blue.", albeit being a claim, is not interesting to the general public and thus should not be fact-checked.

4. Harmfulness: To what extent is the tweet harmful to the society/person(s)/company(s)/product(s)? The purpose of this question is to determine whether the content of the tweet aims to and can negatively affect the society as a whole, a specific person(s), a company(s), a product(s), or could spread rumors about them.[1]

[1] A rumor is a form of a statement whose veracity is not quickly or ever confirmed.

	Train	Dev	Test	Total
Arabic	520	2,536	1,000	4,056
Bulgarian	3,000	350	357	3,707
English	867	53	418	1,338

Table 1: Task 1: Statistics about the dataset.

	Train	Dev	Test	Total
censored	762	93	98	953
uncensored	750	96	91	937
Total	1,512	189	189	1,890

Table 2: Task 2: Statistics about the dataset.

5. Need to Fact-Check: Do you think that a professional fact-checker should verify the claim in the tweet? Not all factual claims are important or worth fact-checking by a professional fact-checker as this is a time-consuming process. For example, claims that could be fact-checked with a very simple search on the Internet probably do not need the attention of a professional fact-checker.

6. Harmful to Society: Is the tweet harmful for the society? The purpose of this question is to judge whether the content of the tweet is could be potentially harmful for the society, e.g., by being weaponized to mislead a large number of people. For example, a tweet might not be harmful because it is a joke, or it might be harmful because it spreads panic, rumors or conspiracy theories, promotes bad cures, or is xenophobic, racist, or hateful.

7. Requires Attention: Do you think that this tweet should get the attention of government entities? A variety of tweets might end up in this category, e.g., such blaming the authorities, calling for action, offering advice, discussing actions taken or possible cures, asking important questions (e.g., "Will COVID-19 disappear in the summer?"), etc.

Data: For this task, the dataset covers three different languages (Arabic, Bulgarian, and English), annotated with yes/no answers to the above questions. More details about the data collection and the annotation process, as well as statistics about the corpus can be found in (Alam et al., 2021b, 2020), where an earlier (and much smaller) version of the corpus is described. We annotated additional tweets for Arabic and Bulgarian for the shared task using the same annotation schema. Table 1 shows the distribution of the examples in the training, development and test sets for the three languages. Note that, we have more data for Arabic and Bulgarian than for English.

3.2 Task 2: Censorship Detection

Task Setup: For this task, we deal with a particular type of censorship – when a post gets removed from a social media platform semi-automatically based on its content. The goal is to predict which posts on Sina Weibo, a Chinese microblogging platform, will get removed from the platform, and which posts will remain on the website.

Data: Tracking censorship topics on Sina Weibo is a challenging task due to the transient nature of censored posts and the scarcity of censored data from well-known sources such as FreeWeibo[2] and WeiboScope[3]. The most straightforward way to collect data from a social media platform is to make use of its API. However, Sina Weibo imposes various restrictions on the use of its API[4] such as restricted access to certain endpoints and restricted number of posts returned per request. Above all, their API does not provide any endpoint that allows easy and efficient collection of the target data (posts that contain sensitive keywords). Therefore, Ng et al. (2019) and Ng et al. (2020) developed an alternative method to track censorship for our purposes. The reader is referred to the original articles to learn more details about the data collection. In a nutshell, the dataset contains censored and uncensored tweets, and it includes no images, no hyperlinks, no re-blogged content, and no duplicates.

For the present shared task 2, we use the balanced dataset described in (Ng et al., 2020) and (Ng et al., 2019). The data is collected across ten topics for a period of four months: from August 29, 2018 till December 29, 2018. Table 2 summarizes the datasets in terms of number of censored and uncensored tweets in the training, development, and testing sets, while Table 3 shows the main topics covered by the dataset.

[2] http://freeweibo.com
[3] http://weiboscope.jmsc.hku.hk
[4] http://open.weibo.com/wiki/API文档/en

Topic	Censored	Uncensored
cultural revolution	55	60
human rights	53	67
family planning	15	25
censorship & propaganda	32	54
democracy	119	107
patriotism	70	105
China	186	194
Trump	320	244
Meng Wanzhou	55	76
kindergarten abuse	48	5
Total	953	937

Table 3: Task 2: Topics featured in the dataset.

4 Task Organization

In this section, we describe the overall task organization, phases, and evaluation measures.

4.1 Task Phases

We ran the shared tasks in two phases:

Development Phase In the first phase, only training and development data were made available, and no gold labels were provided for the latter. The participants competed against each other to achieve the best performance on the development set.

Test Phase In the second phase, the test set (unlabeled input only) was released, and the participants were given a few days to submit their predictions.

4.2 Evaluation Measures

The official evaluation measure for task 1 was the average of the weighted F1 scores for each of the seven questions; for task 2, it was accuracy.

5 Evaluation Results for Task 1

Below, we describe the baselines, the evaluation results, and the best systems for each language.

5.1 Baselines

The baselines for Task 1 are (i) majority class, (ii) ngram, and (iii) random. The performance of these baselines on the official test set is shown in Tables 4, 5, and 6.

5.2 Results and Best Systems

The results on the official test set for English, Arabic, and Bulgarian are reported in Tables 4, 5, and 6, respectively. We can see that most participants managed to beat all baselines by a margin.

Below, we give a brief summary of the best performing systems for each language.

The English Winner: Team TOKOFOU (Tziafas et al., 2021) performed best for English. They gathered six BERT-based models pre-trained in relevant domains (e.g., Twitter and COVID-themed data) or fine-tuned on tasks, similar to the shared task's topic (e.g., hate speech and sarcasm detection). They fine-tuned each of these models on the task 1 training data, projecting a label from the sequence classification token for each of the seven questions in parallel. After model selection on the basis of development set F1 performance, they combined the models in a majority-class ensemble.

The Arabic Winner: Team R00 had the best performing system for Arabic. They used an ensemble of the follwoing fine-tuned Arabic transformers: AraBERT (Antoun et al., 2020), Asafaya-BERT (Safaya et al., 2020), ARBERT. In addition, they also experimented with MARBERT (Abdul-Mageed et al., 2020).

The Bulgarian Winner: We did not receive a submission for the best performing team for Bulgarian. The second best team, HunterSpeechLab (Panda and Levitan, 2021), explored the cross-lingual generalization ability of multitask models trained from scratch (logistic regression, transformer encoder) and pre-trained models (English BERT, and mBERT) for deception detection.

5.3 Summary of All Systems

DamascusTeam (Hussein et al., 2021) used a two-step pipeline, where the first step involves a series of pre-processing procedures to transform Twitter jargon, including emojis and emoticons, into plain text. In the second step, a version of AraBERT is fine-tuned and used to classify the tweets. Their system was ranked 5th for Arabic.

Team dunder_mifflin (Suhane and Kowshik, 2021) built a multi-output model using task-wise multi-head attention for inter-task information aggregation. This was built on top of the representations obtained from RoBERTa. To tackle the small size of the dataset, they used back-translation for data augmentation. Their loss function was weighted for each output, in accordance with the distribution of the labels for that output. They were the runners-up in the English subtask with a mean F1-score of 0.891 on the test set, without the use of any task-specific embeddings or ensembles.

Rank	Team	F1	P	R	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	TOKOFOU	0.897	0.907	0.896	0.835	0.913	0.978	0.873	0.882	0.908	0.889
2	dunder_mifflin	0.891	0.907	0.878	0.807	0.923	0.966	0.868	0.852	0.940	0.884
3	NARNIA	0.881	0.900	0.879	0.831	0.925	0.976	0.822	0.854	0.909	0.849
4	InfoMiner	0.864	0.897	0.848	0.819	0.886	0.946	0.841	0.803	0.884	0.867
5	advex	0.858	0.882	0.864	0.784	0.927	0.987	0.858	0.703	0.878	0.866
6	LangResearchLabNC	0.856	0.909	0.827	0.842	0.873	0.914	0.829	0.792	0.894	0.849
	majority_baseline	0.830	0.786	0.883	0.612	0.927	1.000	0.770	0.807	0.873	0.821
	ngram_baseline	0.828	0.819	0.868	0.647	0.904	0.992	0.761	0.800	0.873	0.821
7	HunterSpeechLab	0.736	0.874	0.684	0.738	0.822	0.824	0.744	0.426	0.878	0.720
8	spotlight	0.729	0.907	0.676	0.813	0.822	0.217	0.764	0.701	0.905	0.877
	random_baseline	0.496	0.797	0.389	0.552	0.480	0.457	0.473	0.423	0.563	0.526

Table 4: Task 1, English: Evaluation results. For Q1 to Q7, the results are in terms of weighted F1 score.

Rank	Team	F1	P	R	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	R00	0.781	0.842	0.763	0.843	0.762	0.890	0.799	0.596	0.912	0.663
	iCompass	0.748	0.784	0.737	0.797	0.746	0.881	0.796	0.544	0.885	0.585 2
	HunterSpeechLab	0.741	0.804	0.700	0.797	0.729	0.878	0.731	0.500	0.861	0.690 3
	advex	0.728	0.809	0.753	0.788	0.821	0.981	0.859	0.573	0.866	0.205 4
	InfoMiner	0.707	0.837	0.639	0.852	0.704	0.774	0.743	0.593	0.698	0.588
	ngram_baseline	0.697	0.741	0.716	0.410	0.762	0.950	0.767	0.553	0.856	0.579
5	DamascusTeam	0.664	0.783	0.677	0.169	0.754	0.915	0.783	0.583	0.857	0.589
	majority_baseline	0.663	0.608	0.751	0.152	0.786	0.981	0.814	0.475	0.857	0.579
6	spotlight	0.661	0.805	0.632	0.843	0.703	0.792	0.647	0.194	0.828	0.620
	random_baseline	0.496	0.719	0.412	0.510	0.444	0.487	0.442	0.476	0.584	0.533

Table 5: Task 1, Arabic: Evaluation results. For Q1 to Q7, the results are in terms of weighted F1 score (The team iCompass submitted their system after the deadline, and thus we rank them with a).

Rank	Team	F1	P	R	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	advex	0.837	0.860	0.861	0.887	0.955	0.980	0.834	0.819	0.678	0.706
2	HunterSpeechLab	0.817	0.819	0.837	0.937	0.943	0.968	0.835	0.748	0.605	0.686
	majority_baseline	0.792	0.742	0.855	0.876	0.951	0.986	0.822	0.672	0.606	0.630
	ngram_baseline	0.778	0.790	0.808	0.909	0.919	0.949	0.803	0.631	0.606	0.630
3	spotlight	0.686	0.844	0.648	0.832	0.926	0.336	0.669	0.687	0.650	0.700
4	InfoMiner	0.578	0.826	0.505	0.786	0.749	0.419	0.599	0.556	0.303	0.631
	random_baseline	0.496	0.768	0.400	0.594	0.502	0.470	0.480	0.399	0.498	0.528

Table 6: Task 1, Bulgarian: Evaluation results. For Q1 to Q7 results are in terms of weighted F1 score.

Team HunterSpeechLab (Panda and Levitan, 2021) participated in all three languages. They explored the cross-lingual generalization ability of multitask models trained from scratch (logistic regression, transformers) and pre-trained models (English BERT, mBERT) for deception detection. They were 2nd for Arabic and Bulgarian.

Team iCompass (Henia and Haddad, 2021) had a late submission for Arabic, and would have ranked 2nd. They used contextualized text representations from ARBERT, MARBERT, AraBERT, Arabic ALBERT and BERT-base-arabic, which they fine-tuned on the training data for task 1. They found that BERT-base-arabic performed best.

Team InfoMiner (Uyangodage et al., 2021) participated in all three subtasks, and were ranked 4th on all three. They used pre-trained transformer models, specifically BERT-base-cased, RoBERTa-base, BERT-multilingual-cased, and AraBERT. They optimized these transformer models for each question separately and used undersampling to deal with the fact that the data is imbalanced.

Team NARNIA (Kumar et al., 2021) experimented with a number of Deep Learning models, including different word embeddings such as Glove and ELMo, among others. They found that the BERTweet model achieved the best overall F1-score of 0.881, securing them the third place on the English subtask.

Team R00 (Qarqaz et al., 2021) had the best performing system for the Arabic subtask. They used an ensemble of neural networks combining a linear layer on top of one out of the following four pre-trained Arabic language models: AraBERT, Asafaya-BERT, ARBERT. In addition, they also experimented with MARBERT.

Team TOKOFOU (Tziafas et al., 2021) participated in English only and theirs was the winning system for that language. They gathered six BERT-based models pre-trained in relevant domains (e.g., Twitter and COVID-themed data) or fine-tuned on tasks, similar to the shared task's topic (e.g., hate speech and sarcasm detection). They fine-tuned each of these models on the task 1 training data, projecting a label from the sequence classification token for each of the seven questions in parallel. After carrying out model selection on the basis of the F1 score on the development set, they combined the models in a majority-class ensemble in order to counteract the small size of the dataset and to ensure robustness.

5.4 Summary of the Approaches

Tables 7, 8 and 9 offer a high-level comparison of the approaches taken by the participating systems for English, Arabic and Bulgarian, respectively (unfortunately, in these comparisons, we miss two systems, which did not submit a system description paper). We can see that across all languages, the participants have used transformer-based models, monolingual or multilingual. In terms of models, SVM and logistic regression were used. Some teams also used ensembles and data augmentation.

Ranks Team	Trans.		Models			Repres.		Misc	
	BERT	RoBERTa	Logistic Regression	SVM	ELMo	GloVe	Ensemble	Under/Over-Sampling	Data Augmentation
1. TOKOFOU									3.
2. dunder_mifflin	¸		¸	¸		¸			
NARNIA									
4. InfoMiner	¸							7.	
HunterSpeechLab									

1 (Tziafas et al., 2021)
2 (Suhane and Kowshik, 2021)
3 (Kumar et al., 2021)
4 (Uyangodage et al., 2021)
7 (Panda and Levitan, 2021)

Table 7: Task 1: Overview of the approaches used by the participating systems for English. =part of the official submission; ¸=considered in internal experiments; Trans. is for Transformers; Repres. is for Representations. References to system description papers are shown below the table.

Ranks Team	Trans.						Models		Misc
	BERT multilingual	AraBERT	Asafaya-BERT	ARBERT	ALBERT	MARBERT	Logistic Regression	Ensemble	Under/Over-Sampling
1. R00	* iCampass								
	¸	¸	¸						
2. HunterSpeechLab									
4. InfoMiner	¸								5.
DamascusTeam									

1 (Qarqaz et al., 2021)
(Henia and Haddad, 2021) 2
(Panda and Levitan, 2021) 4
(Uyangodage et al., 2021) 5
(Hussein et al., 2021)

Table 8: Task 1: Overview of the approaches used by the participating systems for Arabic.

Ranks Team	Trans.	Models	Misc
	BERT multilingual	Logistic Regression	Under/Over-Sampling
2. HunterSpeechLab			
4. InfoMiner			

2 (Panda and Levitan, 2021)
4 (Uyangodage et al., 2021)

Table 9: Task 1: Overview of the approaches used by the participating systems for Bulgarian.

Team	P	R	F1	A
NITK_NLP	c: 0.69 u: 0.61	c: 0.56 u: 0.73	c: 0.62 u: 0.66	0.64
Baseline from (Ng et al., 2020)	c: 0.82 u: 0.76	c: 0.79 u: 0.79	c: 0.80 u: 0.77	0.80
Majority baseline				0.50
Human baseline (Ng et al., 2020)				0.24

Table 10: Task 2: the NITK_NLP team's results. Here: c is censored and u is uncensored.

6 Evaluation Results for Task 2

Below, we report the results for the baselines and for the participating system.

6.1 Baselines

For task 2, we have three baselines as shown in Table 10: a majority class baseline, as before, and two additional baselines described in (Ng et al., 2020). The first additional baseline is a human baseline based on crowdsourcing. The second additional baseline is a multilayer perceptron (MLP) using linguistic features as well as such measuring the complexity of the text, e.g., in terms of its readability, ambiguity, and idiomaticity. These features are motivated by observations that censored texts are typically more negative, more idiomatic, contain more content words and more complex semantic categories. Moreover, censored tweets use more verbs, which indirectly points to the Collective Action Potential. In contrast, uncensored posts are generally more positive, and contain words related to leisure, reward, and money.

6.2 Results

Due to the unorthodox application, and perhaps to the sensitivity of the data, task 2 received only one submission: from team NITK_NLP. The team used a pre-trained XLNet-based Chinese model by Cui et al. (2020), which they fine-tuned for 20 epochs, using the Adam optimizer. The evaluation results for that system are shown in Table 10. We can see that while the system outperformed both the human baseline and the majority class baseline by a large margin, it could not beat the MLP baseline. This suggests that capturing the linguistic fingerprints of censorship might indeed be important, and thus probably should be considered, e.g., in combination with deep contextualized representations from transformers (Ng et al., 2018a, 2019, 2020).

7 Conclusion and Future Work

We have presented the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic in social media (offered in Arabic, Bulgarian, and English) and on censorship detection (offered in Chinese).

In future work, we plan to extend the dataset to cover more examples, e.g., from more recent periods when the attention has shifted from COVID-19 in general to vaccines. We further plan to develop similar datasets for other languages.

Ethical Considerations

While our datasets do not contain personally identifiable information, creating systems for our tasks could face a "dual-use dilemma," as they could be misused by malicious actors. Yet, we believe that the need for replicable and transparent research outweigh concerns about dual-use in our case.

Acknowledgments

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

This research is part of the Tanbih mega-project, developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news," propaganda, and media bias by making users aware of what they are reading.

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

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

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. arXiv/2101.01785.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2021. Mega-COV: A billion-scale dataset of 100+ languages for COVID-19. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL '21, pages 3402–3420.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021a. A survey on multimodal disinformation detection. arXiv/2103.12541.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2021b. Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. In Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '21.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. 2020. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. arXiv/2005.00033.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, OSACT '20, pages 9–15, Marseille, France.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2012. Censorship and deletion practices in Chinese social media. First Monday, 17(3).

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

Kevin Bock, Yair Fax, Kyle Reese, Jasraj Singh, and Dave Levin. 2020. Detecting and evading censorship-in-depth: A case study of Iran's protocol whitelister. In Proceedings of the 10th USENIX Workshop on Free and Open Communications on the Internet, FOCI '20.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. JMIR, 6(2):e19273.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. Sci. Reports, 10(1):1–10.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 657–668.

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

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, ANLP '21, pages 72–81, Kyiv, Ukraine (Virtual).

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis- and disinformation identification. arXiv/2103.00242.

Wassim Henia and Hatem Haddad. 2021. iCompass at NLP4IF-2021–Fighting the COVID-19 infodemic. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

Chaya Hiruncharoenvate, Zhiyuan Lin, and Eric Gilbert. 2015. Algorithmically bypassing censorship on Sina Weibo with nondeterministic homophone substitutions. In Proceedings of the Ninth International Conference on Web and Social Media, ICWSM '15, pages 150–158, Oxford, UK.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.

Ahmad Hussein, Nada Ghneim, and Ammar Joukhadar. 2021. DamascusTeam at NLP4IF2021: Fighting the Arabic COVID-19 Infodemic on Twitter using AraBERT. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

Amir Karami, Morgan Lundy, Frank Webb, Gabrielle Turner-McGrievy, Brooke W McKeever, and Robert

McKeever. 2021. Identifying and analyzing health-related themes in disinformation shared by conservative and liberal Russian trolls on Twitter. Int. J. Environ. Res. Public Health, 18(4):2159.

Sachin Katti, Dina Katabi, and Katarzyna Puchala. 2005. Slicing the onion: Anonymous routing without PKI. Technical report, MIT CSAIL Technical Report 1000.

Gary King, Jennifer Pan, and Margaret E Roberts. 2013. How censorship in China allows government criticism but silences collective expression. American Political Science Review, 107(2):1–18.

Jeffrey Knockel, Masashi Crete-Nishihata, Jason Q. Ng, Adam Senft, and Jedidiah R. Crandall. 2015. Every rose has its thorn: Censorship and surveillance on social video platforms in China. In Proceedings of the 5th USENIX Workshop on Free and Open Communications on the Internet, FOCI '15, Washington, D.C., USA.

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

Ankit Kumar, Naman Jhunjhunwala, Raksha Agarwal, and Niladri Chatterjee. 2021. NARNIA at NLP4IF-2021: Identification of misinformation in COVID-19 tweets using BERTweet. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

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.

Christopher S. Leberknight, Mung Chiang, and Felix Ming Fai Wong. 2012. A taxonomy of censors and anti-censors: Part I: Impacts of internet censorship. International Journal of E-Politics (IJEP), 3(2).

Siu-yau Lee. 2016. Surviving online censorship in China: Three satirical tactics and their impact. The China Quarterly, 228:1061–1080.

Yan Leng, Yujia Zhai, Shaojing Sun, Yifei Wu, Jordan Selzer, Sharon Strover, Hezhao Zhang, Anfan Chen, and Ying Ding. 2021. Misinformation during the COVID-19 outbreak in China: Cultural, social and political entanglements. IEEE Trans. on Big Data, 7(1):69–80.

Dave Levin, Youndo Lee, Luke Valenta, Zhihao Li, Victoria Lai, Cristian Lumezanu, Neil Spring, and Bobby Bhattacharjee. 2015. Alibi routing. In Proceedings of the 2015 ACM Conference of the Special Interest Group on Data Communication, SIGCOMM '15, pages 611–624, London, UK.

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.

Richard J Medford, Sameh N Saleh, Andrew Sumarsono, Trish M Perl, and Christoph U Lehmann. 2020. An "Infodemic": Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. OFID, 7(7).

Azzam Mourad, Ali Srour, Haidar Harmanai, Cathia Jenainati, and Mohamad Arafeh. 2020. Critical impact of social networks infodemic on defeating coronavirus COVID-19 pandemic: Twitter-based study and research directions. IEEE TNSM, 17(4):2145–2155.

Hamdy Mubarak and Sabit Hassan. 2021. ArCorona: Analyzing Arabic tweets in the early days of coronavirus (COVID-19) pandemic. In Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 1–6.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021a. Automated fact-checking for assisting human fact-checkers. arXiv/2103.07769.

Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021b. Detecting abusive language on online platforms: A critical analysis. arXiv/2103.00153.

Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. 2021c. A survey on predicting the factuality and the bias of news media. arXiv/2103.12506.

Kei Yin Ng, Anna Feldman, and Chris Leberknight. 2018a. Detecting censorable content on Sina Weibo: A pilot study. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence, SETN '18, Patras, Greece.

Kei Yin Ng, Anna Feldman, and Jing Peng. 2020. Linguistic fingerprints of internet censorship: the case of Sina Weibo. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI '20, pages 446–453.

Kei Yin Ng, Anna Feldman, Jing Peng, and Chris Leberknight. 2018b. Linguistic characteristics of censorable language on SinaWeibo. In Proceedings of the First Workshop on Natural Language Processing for Internet Freedom, NLP4IF '18, pages 12–22, Santa Fe, New Mexico, USA.

Kei Yin Ng, Anna Feldman, Jing Peng, and Chris Leberknight. 2019. Neural network prediction of censorable language. In Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, pages 40–46, Minneapolis, Minnesota, USA.

Subhadarshi Panda and Sarah Ita Levitan. 2021. Detecting multilingual COVID-19 misinformation on social media via contextualized embeddings. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

Cristina M Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. 2020. COVID-19 infodemic: More retweets for science-based in-formation on coronavirus than for false information. International Sociology, 35(4):377–392.

Ahmed Qarqaz, Dia Abujaber, and Malak A. Abdullah. 2021. R00 at NLP4IF-2021: Fighting COVID-19 infodemic with transformers and more trans-formers. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Free-dom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. SIGSPATIAL Special, 12(1):6–15.

Iris Safaka, Christina Fragouli, and Katerina Argyraki. 2016. Matryoshka: Hiding secret communication in plain sight. In Proceedings of the 6th USENIX Workshop on Free and Open Communications, FOCI '16.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval '20, pages 2054–2059, Barcelona, Spain.

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.

Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. 2020. COVID-19 open source data sets: A comprehensive survey. Applied Intelligence, pages 1–30.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for COVID-19 disinformation categorisation. PLOS ONE, 16(2).

Ayush Suhane and Shreyas Kowshik. 2021. Multi output learning using task wise attention for predicting binary properties of tweets: Shared-task-on-fighting the COVID-19 infodemic. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

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, New Mexico, USA.

Giorgos Tziafas, Konstantinos Kogkalidis, and Tommaso Caselli. 2021. Fighting the COVID-19 infodemic with a holistic BERT ensemble. In Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, NLP4IF@NAACL' 21.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021. Transformers to fight the COVID-19 infodemic. In Proceedings of the Fourth Workshop on Natural Language Processing for Inter-net Freedom: Censorship, Disinformation, and Pro-paganda, NLP4IF@NAACL' 21.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In Proceedings of the Workshop on Online Abuse and Harms, pages 162–172.

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

Zachary Weinberg, Jeffrey Wang, Vinod Yegneswaran, Linda Briesemeister, Steven Cheung, Frank Wang, and Dan Boneh. 2012. StegoTorus: A camouflage proxy for the Tor anonymity system. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12, page 109–120, Raleigh, North Carolina, USA.

Han Zhang and Jennifer Pan. 2019. CASM: A deep-learning approach for identifying collective action events with text and image data from social media. Sociological Methodology, 49(1):1–57.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A multimodal repository for COVID-19 news credibility research. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, page 3205–3212, Galway, Ireland (Virtual Event).

Tao Zhu, David Phipps, Adam Pridgen, Jedidiah R. Crandall, and Dan S. Wallach. 2013. The velocity of censorship: High-fidelity detection of microblog post deletions. In Proceedings of the 22nd USENIX Conference on Security, SEC'13, pages 227–240, Washington, D.C., USA.