SMM4H 2022 Task 2: Dataset for stance and premise detection in tweets about health mandates related to COVID-19

Vera Davydova
Sber AI
Moscow, Russia
veranchos@gmail.com

Elena Tutubalina
Kazan Federal University, Kazan, Russia
Sber AI, Moscow, Russia
AIRI, Moscow, Russia
tutubalinaev@gmail.com

Abstract

This paper is an organizers' report of the competition on argument mining systems dealing with English tweets about COVID-19 health mandates. This competition was held within the framework of the SMM4H 2022 Workshop. During the competition, the participants were offered two subtasks: stance detection and premise classification. We present a manually annotated corpus containing 6,156 short posts from Twitter on three topics related to the COVID-19 pandemic: school closures, stay-at-home orders, and wearing masks. We hope the prepared dataset will support further research on argument mining in the health field.

1 Introduction

Nowadays, people are actively sharing their views on various issues related to the COVID-19 pandemic on social media. For example, users express their attitude towards a quarantine and wearing masks in public places. Some statements are reasoned by arguments, and other statements are just emotional claims. Automated approaches for detecting people's stances and their premises towards health orders related to COVID-19 can help to estimate the level of cooperation within the health mandates announced by the government.

Thereby, since the beginning of the pandemic, new datasets for argument mining in the health field have been created. The first and the largest dataset of Twitter users' stances in the context of the COVID-19 pandemic is COVID-CQ (Mutlu et al., 2020). It contains controversial tweets about the efficacy of hydroxychloroquine as a treatment. Similarly, Wührl and Klinger (2021) presented a dataset for biomedical claim detection in Twitter posts. Miao et al. (2020) created the Lockdown-Tweets – mostly unlabelled tweet dataset, which is related to the lockdown policy in New York State during the pandemic. People's opinions towards health mandates in Germany are discovered in Beck et al. (2021): first, relevant tweets were identified and then the expressed stances were detected.

While most researchers concentrate on the stance detection task, there are far fewer datasets for premise classification (Kotelnikov et al., 2022).

In this work, we aim to fill in this gap and focus not only on stance detection, but also on premise classification. Therefore, we organised a competition on detecting both of these argument structures from English tweets related to COVID-19 health mandates. This competition was carried out as one of the shared tasks during the Social Media Mining for Health Applications (#SMM4H) 2022 Workshop. The SMM4H shared tasks aim to take a community-driven approach to address NLP challenges of utilising social media data for health informatics, including informal, colloquial expressions of clinical concepts, noise, data sparsity, ambiguity, and multilingual posts (Klein et al., 2020), (Magge et al., 2021). In 2022, the seventh iteration of the SMM4H shared tasks, including Task 2 on automatic classification of stance and premise in tweets about health mandates related to COVID-19 (in English), was held (Weissenbacher et al., 2022).

The dataset for this task contains tweets that express views towards three claims/topics: (i) keeping schools closed, (ii) stay-at-home orders, and (iii) wearing a face mask. The main task consists of two sub-tasks: (i) **Task 2a** on stance detection, and (ii) **Task 2b** on premise classification.

Task 2a: stance detection The first task aims to determine the point of view (stance) of the text's author concerning the given claim (e.g., wearing a face mask). The tweets are manually annotated for stance according to three categories: in-favor, against, and neither.

Task 2b: premise classification The second subtask is to predict whether at least one premise/argument is mentioned. A given tweet

Proceedings of the 29th International Conference on Computational Linguistic, pages 216–220
October 12–17, 2022.

is considered as having a premise if it contains a statement that can be used as an argument in a discussion. For instance, the annotator could use it to convince an opponent about the given claim. The tweets are manually annotated for binary classification: participants of this task are required to submit whether each tweet has a premise (1) or not (0).

The main contributions of this work are the following. Firstly, we complemented existing dataset of COVID-19 Stance detection (Glandt et al., 2021) with premise classification labels. Furthermore, guidelines for annotating texts that contain premises are prepared. Secondly, we released the baselines for both subtasks that use RoBERTa architecture. Finally, we compared and analysed the results of the participants of the shared task and proposed steps for further improvement. All the materials can be found on SMM4H Workshop page[1] and on Codalab[2] competition page.

Claim/Topic	**Stance**			**Premise**	
	favor	against	neither	1	0
train set					
face masks	652	324	343	508	811
close school	526	217	307	535	515
home orders	168	333	686	288	899
validation set					
face masks	121	51	36	82	126
close school	91	35	51	80	97
home orders	32	72	111	58	157
test set					
face masks	209	208	260	253	424
close school	215	192	263	294	376
home orders	102	170	381	169	484

Table 1: Summary of statistics of stance and premise classification datasets. The topic on *school closures* and *stay at home orders* has been shortened to *close school* an *home orders*, respectively.

2 Datasets

We provided the participants of the SMM4H 2022 competition the training and validation sets. During the evaluation phase, all participants had five days to submit their predictions for test sets on Codalab for evaluation. The training set included 3,556 tweets, a good mix of three topics: 37%, 33% and 30% of tweets are about face masks, school closures and stay-at-home orders, respectively. The validation and test sets contained 600 and 2,000 tweets, respectively.

As a source of tweets for training and validation sets, we leveraged a COVID-19 stance detection dataset (Glandt et al., 2021) along with stance annotation guidelines and stance labels. Tweets for test sets were collected using 33 keywords such as #OpenSchools, #LockdownNow, #NoMasks. After this, we removed the hashtags to exclude annotation bias toward specific classes. The test set for Task 2a and all three sets for Task 2b were manually annotated. Each tweet was labelled by three *Yandex.Toloka*[3] annotators. We followed annotation guidelines of an argument mining shared task RuArg-2022 (Kotelnikov et al., 2022). Below we highlight some of the key features of our guidelines:

- A statement is evaluated as an argument if it contains a statement that can be used in a dispute to persuade an opponent. For example, *masks help prevent the spread of the disease.* (1)
- It is also necessary to distinguish sentiment (positive and/or negative) from argumentation. For example, *and the fact that Trump did not introduce a suffocating quarantine is well done!* (0)
- The argument should not be a fragment that needs to be thought out. For example, *It is effective if you declare a quarantine.* (0)
- An example of an argument could be such a common sense statement. For example, *in all countries of the world, everyone is wearing masks, but ours... this is not a joke.* (1)
- The position of the author "favor" or "against" should be clear – only under this condition it is possible to detect an argument. The annotator should not think for the author. For example, the author's position on quarantine is unclear in the text *here are the words of my classmate from Annecy, France, from today's Facebook correspondence - "France introduced quarantine, and immediately everyone poured out to barbecue in nature."* (0).

To measure annotation quality on this platform, we mixed raw tweets with control tasks (tweets

[1] https://healthlanguageprocessing.org/smm4h-2022/
[2] https://codalab.lisn.upsaclay.fr/competitions/5067
[3] https://toloka.yandex.ru/

Tweet	Claim/Topic	Stance	Premise
The fact that anti-masking is a thing is a completely terrifying insight into the nature of some beings who look, walk and breathe just like us.	face masks	favor	0
Masks help prevent the spread of the disease. Please, #WEARAMASK	face masks	favor	1
We are now experiencing a surge in the number of infected health care workers, with two deaths already. Prior to Covid19, we were experiencing a shortage and this is worsening with them in quarantine. You can help us by staying safe and staying home.	stay at home orders	favor	1
0.02% chance of dying of #Covid and @GovInslee keeps our state in an “indefinite” lockdown. I’ll take those odd, thanks.	stay at home orders	against	1
I see that @BBCOne are still showing the people on their tandem bikes before programmes. Don’t you lot not know that there is a lockdown and no one can go out right now? #coronavirus	stay at home orders	neither	0
Close the damn schools until there is a vaccine.	school closures	favor	1

Table 2: Examples of tweets annotated for stance and premise classification.

from the validation set with correct responses annotated by both authors). These tasks were used for calculating the Toloker’s percentage of correct responses. Depending on the annotator’s result, the system blocked access to tasks. The internal quality of the control tasks was 0.68. The obtained crowdsourced labels were aggregated into a single label (Dawid and Skene, 1979). Samples of annotated tweets are shown in Table 2.

Table 1 shows statistics of experimental datasets across topics. The training set includes 38% and 25% in-favor and against tweets, respectively. Relative class balance is also present for argumentation: 63% of train tweets contain a premise (1). 34% of tweets in the test set contain a premise; 26% of tweets in the test set are annotated as in-favor. As shown in Table 1, the distribution of classes by topic is different. In particular, the topic about staying-at-home orders contains more “against” tweets than tweets “in-favor”.

3 Experiments

We used F_1 as the main evaluation metric in each of the two subtasks, which is calculated according to the following formula: $F_1 = \frac{1}{n}\sum_{c\in C} F_{1_{rel_c}}$, where C ={“face masks”, “stay at home orders”, “school closures”}, n is the size of C, $F_{1_{rel_c}}$ is macro F_1-score averaged over two classes for each task (in-favor & against classes for stance; 0 & 1 classes for premise).

We used two models as baselines: Random and RoBERTa. Random baseline is rather simplistic: we randomly assigned labels for each tweet in both tasks according to the label distribution. The second baseline leverages RoBERTa architecture (Liu et al., 2019) for multiclass (Task 2a) and binary classification (Task 2b). We fine-tuned pretrained RoBERTa-base model from HuggingFace[4] on our data. The validation set was utilised to select appropriate hyperparameters for the models. For each model, AdamW optimizer (Loshchilov and Hutter, 2019) was used with a learning rate of 4e-5, and gradient clipping with a max norm of 1.0. Each model was trained for 5 epochs, with a mini-batch size of 8 in each iteration and a maximum sequence length of 128.

The results of both baseline models in terms of F_1 scores are described in Table 3. As we can see, RoBERTa-based baseline showed relatively strong results in both set-ups: F_1-score is 0.566 (validation) and 0.45 (test) on the more difficult Stance detection task; and 0.75 (validation) and 0.722 (test) on Premise classification.

[4]https://huggingface.co/roberta-base

Claim/Topic	Stance		Premise	
	Rnd.	RoBERTa	Rnd.	RoBERTa
validation set				
face masks	0.309	0.599	0.423	0.770
close school	0.325	0.513	0.413	0.731
home orders	0.334	0.589	0.406	0.755
All tweets	0.323	0.566	0.414	0.750
test set				
face masks	0.342	0.439	0.506	0.708
close school	0.332	0.345	0.526	0.719
home orders	0.301	0.566	0.521	0.738
All tweets	0.325	0.450	0.518	0.722

Table 3: Macro F_1 scores for both Random (Rnd.) and RoBERTa baselines. The topic on *school closures* and *stay at home orders* has been shortened to *close school* an *home orders*, respectively.

3.1 Official SMM4H 2022 Task 2 Results

We observed a strong interest in Task 2, with 47 participants registered in Codalab. Among these participants, 14 teams submitted their prediction to Codalab for both tasks (15 teams for Task 2b). We summarized their performance in (Weissenbacher et al., 2022). Further, we highlight the key observations. The median F_1 scores of all team's best-performing submissions are 0.55 for Task 2a and 0.65 for Task 2b. The mean F_1 scores of all team's best-performing submissions are 0.49 for Task 2a and 0.57 for Task 2b. The best performance achieved in task 2a is 0.64 F_1 which is 0.19 higher than the RoBERTa baseline model. The three top-performing systems achieved 0.70 F_1 in Task 2b, which is 0.02 less than the baseline model. The majority of teams used COVID-related BERT models. Two teams tried to combine data from both tasks into one unified model. Leading teams on both tasks tried to aggregate claim and tweet texts: the leading team in Task 2a appended the claim text to the tweet, while the second-best team in Task 2a with the highest F_1 in Task 2b proposed a new pre-training task constructed by the tweets and claims similarly to next sentence prediction.

4 Conclusion

In this paper, we have presented the dataset for stance and premise detection in tweets written in English about health mandates related to COVID-19. We hosted a shared task on SMM4H 2022 workshop and released this dataset to the research community. The 15 teams that took part in the task proposed a variety of classification architectures, ranging from just fine-tuning BERT models to multi-task learning on both subtasks and combining tweets with claims. We plan to extend our dataset for future work with a broader set of health-related claims.

Acknowledgements

The work was supported by the Russian Science Foundation [grant number 18-11-00284]. The authors would also like to thank Natalia Loukachevitch for her suggestions on task definition, Dmitry Ustalov and other members of the *Yandex.Toloka* team for providing credits for the crowd-sourced annotation of tweets.

References

Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in German covid-19 social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–13, Online. Association for Computational Linguistics.

Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36, Barcelona, Spain (Online). Association for Computational Linguistics.

Evgeny Kotelnikov, Natalia Loukachevitch, Irina Nikishina, and Alexander Panchenko. 2022. Ruarg-2022: Argument mining evaluation. *arXiv preprint arXiv:2206.09249*.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre-Maduell, Salvador Lima Lopez, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan M Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez, editors. 2021. *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. Association for Computational Linguistics, Mexico City, Mexico.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ece Mutlu, Toktam Oghaz, Jasser Jasser, Ege Tütüncüler, Amirarsalan Rajabi, Aida Tayebi, Ozlem Ozmen, and Ivan Garibay. 2020. A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications smm4h shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop Shared Task*, pages –.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.