Stance Detection in COVID-19 Tweets

Kyle Glandt[♣] **Sarthak Khanal**[♣] **Yingjie Li**[♢] **Doina Caragea**[♣] **Cornelia Caragea**[♢]

[♣]Computer Science, Kansas State University
[♢]Computer Science, University of Illinois at Chicago
{kglandt,sarthakk,dcaragea}@ksu.edu
{yli300,cornelia}@uic.edu

Abstract

The prevalence of the COVID-19 pandemic in day-to-day life has yielded large amounts of stance detection data on social media sites, as users turn to social media to share their views regarding various issues related to the pandemic, e.g. stay at home mandates and wearing face masks when out in public. We set out to make use of this data by collecting the stance expressed by Twitter users, with respect to topics revolving around the pandemic. We annotate a new stance detection dataset, called COVID-19-Stance. Using this newly annotated dataset, we train several established stance detection models to ascertain a baseline performance for this specific task. To further improve the performance, we employ self-training and domain adaptation approaches to take advantage of large amounts of unlabeled data and existing stance detection datasets. The dataset, code, and other resources are available on GitHub.[1]

1 Introduction

We live in unprecedented times caused by a global COVID-19 pandemic, which has forced major changes in our daily lives. Given the developments concerning COVID-19, communities and governments need to take appropriate action to mitigate the effects of the novel coronavirus, which is at the root of the pandemic. For example, states in the United States that have imposed strict social distancing mandates were able to slow the growth of the virus within their communities (Courtemanche et al., 2020). For such measures to work, however, it is important that the public fully adhere to these guidelines and mandates. "Pandemic fatigue," or when people become tired of pandemic mandates and begin to ease in adherence, can lead to resurgences of the novel coronavirus (Feuer and Rattner, 2020). To reduce the spread of COVID-19, it is essential to understand the public's opinion on the various initiatives, such as stay at home orders, wearing a face mask in public, school closures, etc. Understanding how the public feels about these mandates could help health officials better estimate the expected efficacy of their mandates, as well as detect pandemic fatigue before it leads to a serious resurgence of the virus.

In the era of Web 2.0, and especially during a pandemic in which people often resort to online communications, social media platforms provide an astounding amount of data relating to the stance and views held by various populations with respect to a variety of current and important topics. However, the total amount of data that is being generated each second makes it impossible for humans alone to fully make use of them. Fortunately, recent developments in deep learning have yielded state-of-the-art performance in text classification.This makes deep learning an ideal solution for extracting and making sense of the large amounts of data currently in circulation on social media sites.

In particular, given the current events, it is evident that automated approaches for detecting the stance of the population towards targets, such as health mandates related to COVID-19, using Twitter posts, or tweets, can help gauge the level of cooperation with the mandates. Stance detection is a natural language processing (NLP) task in which the goal is for a machine to learn how to automatically determine from text alone an author's *stance*, or perspective/view, towards a controversial topic, or *target*. Research in the area of stance detection has yielded accurate results, especially in the United States politics (Mohammad et al., 2017; Ghosh et al., 2019; Xu et al., 2020). However, research on stance detection for targets relevant to COVID-19 health mandates lags behind, due to the

[1] https://github.com/kglandt/stance-detection-in-covid-19-tweets

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1596–1611
August 1–6, 2021.

Tweet	Target	Stance	Opinion	Sentiment
Idc what you say, you're selfish if you refuse to wear a mask. This shouldn't be political. #MaskUp	Wearing a Face Mask	In Favor	Explicit	Negative
That video, my god. I'm as progressive as the next person and I dearly hope Trump will lose, but I can't remember the last time I watched such a cynical, fear-mongering piece of propaganda. Keeping schools closed will be devastating for our most vulnerable children.	Keeping Schools Closed	Against	Explicit	Negative
I believe in SCIENCE. I wear a mask for YOUR PROTECTION. #JoeBidenForPresident2020	Anthony S. Fauci, M.D.	In Favor	Implicit	Positive
"@realDonaldTrump Also Death Rate down BIG TIME! America is ready for Business!"	Stay at Home Orders	Against	Implicit	Positive

Table 1: Examples of tweet/target pairs from the COVID-19-Stance dataset, manually annotated with respect to user's *stance* towards the target, the way stance *opinion* was expressed, and the overall *sentiment* of the tweet.

recency of the pandemic and a lack of benchmark datasets. We set out to address this problem by constructing a COVID-19 stance detection dataset (called COVID-19-Stance), which includes tweets that express views towards four targets, specifically "Anthony S. Fauci, M.D.", "Keeping Schools Closed", "Stay at Home Orders", and "Wearing a Face Mask." This is a challenging task, which is related but different from sentiment analysis. A tweet may express support for a target, while using a negative language, and expressing a negative *sentiment* overall. Furthermore, the *opinion* expressed in a tweet may not be *explicitly* towards the target of interest, while the stance can be *implicitly* inferred. Some examples of tweet/target pairs labeled with respect to *stance*, target of *opinion* and *sentiment* are shown in Table 1 to illustrate the above mentioned challenges.

To address the stance detection task, carefully designed approaches are needed to extract language patterns informative with respect to stance. We provide a comprehensive set of baseline results for the newly constructed COVID-19-Stance dataset, including results with established supervised baselines for stance detection tasks, and also baselines that employ approaches for handling small amounts of labeled data, including self-training and domain adaptation approaches. In summary, the contributions of this work are as follows:

- We construct a COVID-19-Stance dataset that consists of 6,133 tweets covering user's stance towards four targets relevant to COVID-19 health mandates. The tweets are manually annotated for stance according to three categories: *in-favor*, *against*, and *neither*.
- We establish baseline results using state-of-the-art supervised stance detection models, including transformer-based models.
- We also establish baselines for self-training and domain adaptation approaches that use unlabeled data from the current task, or labeled data from a related task, to complement for limited labeled data for the current task.

2 Related Work

We discuss related work in terms of existing datasets and approaches for stance detection.

2.1 Stance Detection Datasets

Recent work on stance detection in social media data has been facilitated by Mohammad et al. (2016, 2017), who constructed a manually annotated stance detection dataset, shared publicly as SemEval2016 Task 6. The dataset was based on tweets about United States politics, collected during the lead up to the United States 2016 presidential election. Given a set of politics-relevant targets (e.g., politicians, feminism, climate change), the initial selection of tweets to be included in the dataset was done using "query hashtags", which are Twitter hashtags within a manually curated shortlist that had been observed to correlate stances and targets on Twitter. Subsequently, tweet/target pairs were annotated by CrowdFlower[2] workers, who were provided with a generic, but detailed questionnaire regarding the stance of a tweet's author toward a target, as well as the sentiment of the tweet (Mohammad et al., 2016, 2017).

Several other datasets for stance detection have become available in the last few years, including a large dataset (containing approximately 50,000 tweets) focused on the stance towards financial transactions that involve mergers and acquisition (Conforti et al., 2020), a dataset for identifying the stance in Twitter replies and quotes (Villa-Cox

[2] http://www.crowdflower.com/

et al., 2020), datasets in languages different from English (Hercig et al., 2017; Vychegzhanin and Kotelnikov, 2019; Evrard et al., 2020), and multilingual datasets (Zotova et al., 2020; Vamvas and Sennrich, 2020; Lai et al., 2020).

Furthermore, the global prevalence and impact of the COVID-19 pandemic has led to the quick development, concurrently with our work, of several COVID-19 stance-related Twitter datasets (Mutlu et al., 2020; Miao et al., 2020; Hossain et al., 2020). Mutlu et al. (2020) published a dataset of approximately 14,000 tweets (called COVID-CQ), which were manually annotated with respect to the author's stance regarding the use of hydroxychloroquine in the treatment of COVID-19 patients. Miao et al. (2020) constructed a dataset focused on author's stance towards lockdown regulations in New York City. The authors used keywords related to "lockdown" and "New York City" and extracted approximately 31,000 relevant tweets from a large COVID-19 tweet dataset published by Chen et al. (2020). They manually annotated 1629 tweet with respect to stance, while the remaining tweets were used as unlabeled.

Our dataset construction procedure is similar to the one followed by Miao et al. (2020), but we label data for four targets using global English tweets, as opposed to Miao et al. (2020) who label data for just one target ("lockdown") in one location ("New York City").

2.2 Stance Detection Approaches

In terms of approaches used for stance detection, strong baseline results based on support vector machines (SVM) with manually engineered features were provided for the SemEval2016 Task 6 by Mohammad et al. (2016, 2017). Deep learning approaches used in SemEval2016 Task 6 included recurrent neural networks (RNNs) (Zarrella and Marsh, 2016) and convolutional neural networks (CNNs) (Vijayaraghavan et al., 2016; Wei et al., 2016). Such approaches used the tweets as input, but did not use any target-specific information, and did not outperform the SVM baselines. Later approaches were provided with both target and tweet representations as input, and employed RNNs and/or CNNs, together with the attention mechanism (Augenstein et al., 2016; Du et al., 2017; Zhou and Cristea, 2017; Sun et al., 2018; Siddiqua et al., 2019) to improve the performance of the SVM baselines.

Given the dominance of transformers (Vaswani et al., 2017), especially bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), in NLP tasks, some recent works (Slovikovskaya and Attardi, 2020; Li and Caragea, 2021; Ghosh et al., 2019) have focused on investigating the use of BERT models for stance detection. For example, Ghosh et al. (2019) explored the reproducibility of approaches for stance detection and compared them to BERT. They found BERT to be the best model overall for stance detection on the SemEval2016 Task 6. Li and Caragea (2021) also explored BERT based models with data augmentation and found BERT to be a powerful model for stance detection. Thus, we have selected BERT as a strong baseline for our paper.

Several works have shown that auxiliary information, such as sentiment and emotion information, or the subjective/objective nature of a text (provided as additional inputs or presented in the form of auxiliary tasks in a multi-task framework), can help improve the performance obtained from the tweet/target information alone (Mohammad et al., 2017; Sun et al., 2019; Li and Caragea, 2019; Hosseinia et al., 2020; Xu et al., 2020). Other approaches to improve the performance, especially when the amount of labeled data for the task of interest is small, include weak supervision (Wei et al., 2019) and knowledge distillation (Miao et al., 2020); transfer learning through distant supervision (Zarrella and Marsh, 2016) or pre-trained models (Ebner et al., 2019; Hosseinia et al., 2020); and domain adaptation from a source task to the target task (Xu et al., 2018, 2020).

In particular, the Dual-view Adaptation Network (DAN) (Xu et al., 2020) learns to predict the stance of a tweet by combining the subjective and objective views/representations of the tweet, while also learning to adapt them across domains. We use an adaptation of the DAN model as a strong baseline in this work. Most relevant to our work on COVID-19-Stance, Miao et al. (2020) compared a supervised in-domain BERT model trained and tested on "lockdown" tweets, with cross-domain models, and knowledge distillation variants. The results showed significantly improved performance for the knowledge distillation variants, and emphasized the importance of having a small amount of data for the task of interest (as a better alternative to zero-shot learning). Similar to Miao et al. (2020), we also use BERT together with knowledge distillation/self-training as a strong baseline.

Target	In-favor	Against
Anthony S. Fauci, M.D.	IStandWithFauci, FauciIsAHero, FauciHero, FireTrumpNotFauci, SaveFauci, IstandWithDrFauci, ThankYouDrFauci, StandWithFauci, ListenToFauci, DrFauciIsANationalHero, ImWithFauci, TrustFauci, LetFauciLead	FireFauci, FauciTheFraud, FauciFraud, FireFauciNow, FraudFauci
Keeping Schools Closed	CloseTheSchools, SchoolsMustShutdown, SaveOurSchools, NotMyChild, SchoolsMustShutdown, CloseTheSchools	RightToLearn, OpenSchools, SchoolReopening, ReopeningSchools
Stay at Home Orders	SaferAtHome, LockdownNow, StayAtHomeSaveLives, StaySafeStayHome	ReopenAmericaNow, ReOpenAmerica, EndTheShutdown, endthelockdown, OPENAMERICANOW, NoToLockdown
Wearing a Face Mask	MasksSaveLives, WearAMaskSaveALife, MaskMoaners, WearAMaskPlease, CoverYourFace, MasksOn, WearADamnMask	MasksOff, SayNoToMasks, NoMasks, NoMask, masksdontwork, MasksOff, MasksOffAmerica, NoMaskOnMe

Table 2: In favor and against query hashtags for each target

3 COVID-19-Stance Dataset

The recency of the COVID-19 pandemic means there was no established stance detection dataset for this broader topic, when we began our research. Therefore, we set out to construct our own dataset, called COVID-19-Stance, by following the methodology introduced by Mohammad et al. (2016, 2017), which is generic and applicable for any controversial topic discussed on Twitter.

Data collection. We began crawling Twitter, using the Twitter Streaming API, on February 27th, 2020. We collected tweets that contained general keywords pertaining to the novel coronavirus (e.g. "coronavirus", "covid-19", "corona virus", "#covid19", etc.). As new hashtags emerged, we iteratively added additional, more specific keywords to the search (e.g., "#lockdown", "stay at home", "#socialdistancing", "#washhands", etc.). We continued crawling until August 20th, 2020. The full list of keywords that was used over this time period is provided in Appendix A. We only stored original tweets (not a retweet or quoted tweet) that contained no hyperlinks, and ended up collecting a grant total of 30,331,993 tweets.

Target selection. After being able to analyze the initial tweets, and following the developments of the COVID-19 events, we began to identify controversial topics that arose as the virus continued its spread in the United States (US). Four topics that we found to be among the most prevalent in our collection of tweets, and are understood by a large number of people in the US, were "Stay at Home Orders", "Wearing a Face Mask", "Keeping Schools Closed", and "Anthony S. Fauci, M.D.".

Data selection. Similar to Mohammad et al. (2016), we identified query hashtags to encompass the four main targets/topics selected, and began to collect and organize the tweets according to topic and likely labels. For example, if "#FireFauci" is contained within a tweet, it is likely that the author of that tweet is posting information indicating they do not support the current director of the National Institute of Allergy and Infectious Diseases (NIAID), Anthony S. Fauci, M.D. For each of the four selected targets, we identified two types of query hashtags, specifically, "in-favor" hashtags and "against" hashtags (stance-neutral hashtags were very rare). The exact query hashtags identified for each target are shown in Table 2. Using the "in-favor" and "against" query hashtags, we selected a "noisy stance set" of tweets for each target, as shown in Table 3. Out of the total number of tweets corresponding to a target, we further selected a relatively balanced (in terms of *in-favor* and *against* noisy labels) dataset to be manually labeled, and another relatively balanced dataset of tweets to be used as unlabeled in the self-training approach. The exact number of tweets *to-label* and to be used *unlabeled* are shown in Table 4.

Target	#In-favor	#Against
Anthony S. Fauci, M.D.	2,417	6,641
Keeping Schools Closed	5,345	5,665
Stay at Home Orders	8,437	5,323
Wearing a Face Mask	27,600	12,064
All	43,799	29,693

Table 3: The number of tweets selected using "in-favor" and "against" hashtags for each target.

Data Annotation. Although query hashtags are great for selecting likely relevant tweets, they are noisy and not reliable enough to accurately identify the stance towards a target for a tweet (see Table 5 for some examples illustrating this point). There-

Target	# to-label	# unlabeled
Anthony S. Fauci, M.D.	2,085	2,443
Keeping Schools Closed	1,479	2,703
Stay at Home Orders	1,717	15,488
Wearing a Face Mask	1,921	9,006
All	7,122	29,640

Table 4: The number of tweets selected to be labeled (#to-label) and the number of tweets to be used as unlabeled in self-training (#unlabeled) for each target.

fore we used Amazon Mechanical Turk (AMT) to enlist the help of gig workers to analyze and label our collection of 7,122 tweets selected to be labeled (the exact number of tweets for each target is shown in Table 4). We removed the hashtags that appeared at the end of a tweet to exclude obvious cues, without making the tweet syntactically ambiguous. This increases the chance that our collection contains tweets that do not explicitly mention the target, and potentially some tweets with neutral stance towards the target. Each tweet was labeled by three annotators. At one time, each annotator was shown a page with a tweet and a target, and asked to answer a questionnaire designed and detailed by Mohammad et al. (2017). The questionnaire, shown in Appendix B, contains detailed questions and multi-choice answers that allow us to annotate each tweet with respect to three criteria:

1. the *stance* of the tweet's author/user towards the given target: *in favor*, *against* or *neither*;

2. the way the *opinion* is expressed, which captures whether the text of the tweet reveals the stance *explicitly*, *implicitly*, or *neither*;

3. the *sentiment* of the tweet, which essentially captures the language used in the tweet: *positive*, *negative*, *both*, *sarcasm*, or *neither*.

Our final COVID-19-Stance dataset contains only tweets for which at least two out of the three annotators agreed on the stance category. The Cohen's Kappa scores that we obtained for inter-annotator agreement for the final dataset were 0.82 for stance, 0.83 for target of opinion, and 0.60 for sentiment. According to (Cohen, 1960), the scores for stance and target of represent almost perfect agreement, while the score for sentiment shows substantial agreement. Table 1 shows several examples of annotated tweets in our dataset.

Dataset statistics. The number of tweets for each target and the stance distribution for each target are shown in Table 6. The number of tweets for

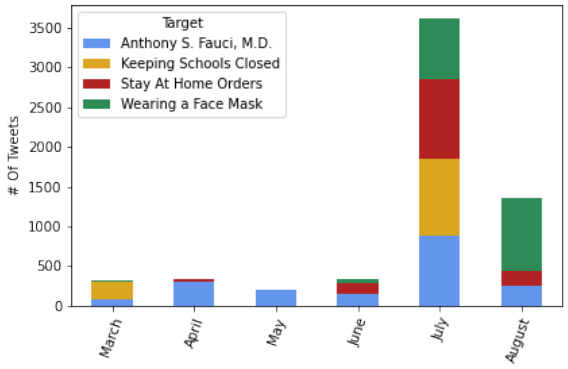

Figure 1: The number of tweets by target over the March to August 2020 months.

each target over the months when data was crawled is graphically displayed in Figure 1, which shows that a large number of the tweets in our dataset were posted in July 2020. The distribution of the type of *opinion* is shown in Tables 7 and 8, for each target and each stance, respectively. Similarly, the distribution of the *sentiment* (or tweet language) is shown in Tables 9 and 10, for each target and each stance, respectively. As can be seen from these tables, our dataset contains a good mix of *in-favor*, *against* and *neutral* categories, and also a good mix of tweets with *implicit* and *explicit* opinion towards the target. However, the sentiment is generally *negative* or in the *other* category (which includes both positive and negative, sarcastic language and neither). Together, these characteristics make our task both realistic and challenging. While we only use the stance label in this work, the other labels will be explored in future works, as auxiliary information potentially useful for stance detection.

Benchmark subsets. To enable progress on COVID-19 stance detection, and facilitate comparisons between models developed for this task, we randomly split our COVID-19-Stance dataset (using stratified sampling) into training (`Train`), development (`Val`) and test (`Test`) subsets, respectively. We used the training subset to train our models, the development to select hyperparameters and the test to evaluate the final performance of the models. Statistics for the dataset in terms of number of tweets in the `Train`, `Test` and `Val` subsets, respectively, are shown in Table 11.

4 Baseline Models

Having described our COVID-19-Stance dataset, we now briefly review several models that we use to establish baseline results on this dataset.

Tweet	Query Tag
125 days. @DeanObeidallah #IcantBreathe #BlackLivesMatter #WednesdayMotivation #Pride #WearAMask #StayHome #VoteByMail #ImRidenWithBiden #Joe2020 #JoeBiden #LockHimUp #TheResistance #StrongerTogether #EqualityForAll #MakeItCount #Enough #LoveIsLove #NeverAgain #LoveWins	WearAMask
I will NOT #WearAMask because the #government says I have to wear one. You #WearADamnMask if you want to. I will not subjugate myself to their Unconstitutional rules. #Idonotcomply #IDoNotConsent	WearADamnMask
"I will not be chipped. I will not be tracked" the DEVIL coronavirus covid 19,all governments will control billions of people, 198 countries!! #illegal #Puppets #NoFacemask #coronavirus #covid19 #pandemic #WearAMask #WearADamnMask #NoFaceMask	WearADamnMask

Table 5: Examples of tweets where the query tags are not reliable silver labels for the Wearing A Face Mask target.

		Distribution of Stances (%)		
Target	**# Total**	**In-favor**	**Against**	**Neither**
Anthony S. Fauci, M.D.	1864	26.39	32.73	40.88
Keeping Schools Closed	1190	51.68	21.01	27.31
Stay at Home Orders	1372	13.85	29.15	57.00
Wearing a Face Mask	1707	40.60	39.13	20.27
All	6133	32.45	31.44	36.12

Table 6: The distribution of stances in the dataset.

	Opinion Towards Target (%)		
Target	**Explicit**	**Implicit**	**Neither**
Anthony S. Fauci, M.D.	44.74	48.34	6.92
Keeping Schools Closed	69.66	26.39	3.95
Stay at Home Orders	23.18	50.36	26.46
Wearing a Face Mask	74.17	22.61	3.22
All	52.94	37.37	9.69

Table 7: The distribution of opinion for each target.

	Opinion Towards Target (%)		
Stance	**Explicit**	**Implicit**	**Neither**
Favor	81.61	17.64	0.75
Against	79.25	20.49	0.26
Neither	4.29	69.80	25.91

Table 8: The distribution of opinion for each stance.

	Sentiment of Tweet (%)		
Target	**Positive**	**Negative**	**Other**
Anthony S. Fauci, M.D.	9.33	69.85	20.82
Keeping Schools Closed	14.62	71.51	13.87
Stay at Home Orders	19.17	46.43	34.40
Wearing a Face Mask	13.24	73.87	12.89
All	13.65	66.05	20.30

Table 9: The distribution of sentiment for each target.

4.1 Supervised Baseline Models

To get a baseline understanding of how established stance detection networks perform on our dataset, we used the following models:

- **BiLSTM**: Bi-Directional Long Short Term Memory Networks (Schuster and Paliwal, 1997) take tweets as input, and are trained to predict the stance towards a target, without explicitly using the target information.
- **Kim-CNN**: Convolutional Neural Networks for text, proposed by Kim (2014), are also provided with tweets as input, and trained to predict the stance towards a target, without explicitly using the target information.
- **TAN**: Target-specific Attention Networks (Du et al., 2017) represent an attention-based BiLSTM model that identifies features specific to the target of interest, by explicitly incorporating the target information.
- **ATGRU**: The Bi-Directional Gated Recurrent Unit Network with Token-Level Attention Mechanism (Zhou and Cristea, 2017) is an attention-based Bi-GRU model that also uses the target information explicitly, and identifies specific target features using the attention.
- **GCAE**: The Gated Convolutional Network with Aspect Embedding (Xue and Li, 2018) is based on a CNN model. In addition to tweets, it also has information about the target, and uses a gating mechanism to block target-unrelated information.
- **BERT**: Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) represent language models that are pre-trained on a large unlabeled corpus to encode sentences and their tokens into dense vector representations. We used the pre-trained COVID-Twitter-BERT model[3] (Müller et al., 2020).

4.2 Self-training Baseline

Given that a large amount of unlabeled data is available for each target included in our COVID-19-Stance dataset, we explored the use of a self-training approach that can make use of unlabeled data, as described below:

[3] https://huggingface.co/digitalepidemiologylab/covid-twitter-bert

	Sentiment of Tweet (%)		
Stance	**Positive**	**Negative**	**Other**
Favor	22.36	62.16	15.48
Against	5.45	87.97	6.59
Neither	12.96	50.47	36.57

Table 10: The distribution of sentiment for each stance.

Target	**# Train**	**# Val**	**# Test**
Anthony S. Fauci, M.D.	1464	200	200
Keeping Schools Closed	790	200	200
Stay at Home Orders	972	200	200
Wearing a Face Mask	1307	200	200

Table 11: The number of tweets for the training (Train), validation (Val) and Test (Test) subsets per target.

- **BERT-NS**: Self-training with Noisy Student (Xie et al., 2020) is a semi-supervised learning approach that employs self-training and knowledge distillation (Hinton et al., 2015) to improve the performance of a teacher model using unlabeled data. More specifically, a teacher is originally trained from the available labeled data, and is used to predict pseudo-labels for the unlabeled data. Subsequently, a noisy student model is trained using the labeled and pseudo-labeled data. By replacing the teacher with the student, the process can be iterated several times. In our work, we performed just one iteration. Both the teacher and the student models were COVID-Twitter-BERT, with a softmax layer at the top.

4.3 Domain Adaptation Baseline

To understand the benefits of using a prior stance detection dataset, in addition to the dataset we constructed, we experimented with a domain adaptation model, as described below:

- **BERT-DAN**: Dual-view Attention Networks (Xu et al., 2020) capture explicitly subjective and objective information contained in tweets, and also enable the use of labeled data for a prior, related task to train a model for a current task of interest. The original DAN model proposed by Xu et al. (2020) makes use of BiLSTM networks and domain adversarial networks to learn the subjective and objective representations and make them domain invariant. At the same time, DAN learns to predict the stance using labeled data from the prior task (under the assumption that no labeled data is available for the task of interest). Compared to the original DAN model, we replaced the BiLSTM networks with pre-trained COVID-Twitter-BERT models, and trained the network to predict the stance using both labeled data from the prior task and from the current task. The prior data was the whole SemEval2016 Task 6 data.

5 Experimental Setup

5.1 Implementation Details

Data Pre-processing Before the tweets in our dataset were used for training, they were pre-processed and transformed to embedded tensors. For every tweet in the dataset, we removed any emojis, URLs, and reserved words. We then used the pre-trained COVID-Twitter-BERT to tokenize and embed each tweet, truncating the sequence length to 128 as needed.

Hyperparameters. The validation set was used to determine generally good hyperparameters for the models. For each non-BERT supervised model, Adam optimizer was used with a learning rate of $1e^{-5}$, weight decay of $4e^{-5}$, and gradient clipping with a max norm of 4.0. Each model was trained for 120 epochs, with a mini-batch size of 16 in each iteration. A dropout of 0.5 was used for each network. Other specific hyper-parameters for each network are shown below:

- **RNN Networks**: BiLSTM, ATGRU, and TAN each had a hidden LSTM dimension of 512 with a dropout of 0.2.

- **CNN Networks**: GCAE and Kim-CNN both used filters of width 2, 3, 4, and 5. For each filter width, there were 25 feature maps. Following the convolutional layers was a linear classifier with a hidden dimension of 128.

- **BERT**: This model was initialized with the pre-trained COVID-Twitter-BERT model. It was optimized with AdamW with a learning rate of $1e^{-5}$ over the course of 10 epochs, with 15 warmup steps.

- **BERT-NS:** The implementation of the student model is exactly the same as that of the supervised BERT. The teacher and the student models are set up in the same manner, except that the teacher has no dropout.

- **BERT-DAN:** The formation functions are the same as those of the supervised BERT model,

		BiLSTM	Kim-CNN	TAN	ATGRU	GCAE	BERT	BERT-NS	BERT-DAN
Target: Anthony S. Fauci, M.D.									
	Acc	0.638	0.633	0.588	0.635	0.652	0.817	0.820	**0.830**
	Pr	0.639	0.685	0.558	0.640	0.661	0.816	0.821	**0.833**
	Re	0.631	0.612	0.564	0.613	0.634	0.830	0.823	**0.839**
Avg.	F1	0.630	0.604	0.547	0.612	0.640	0.818	0.821	**0.832**
Target: Keeping Schools Closed									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.627	0.625	0.598	0.590	0.588	0.772	**0.780**	0.758
	Pr	0.570	0.549	0.545	0.548	0.528	0.765	**0.773**	0.748
	Re	0.545	0.509	0.532	0.528	0.488	**0.761**	0.743	0.702
Avg.	F1	0.548	0.495	0.534	0.527	0.490	**0.755**	0.753	0.717
Target: Stay At Home Orders									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.735	0.703	0.695	0.682	0.738	**0.843**	0.832	0.833
	Pr	0.679	0.552	0.523	0.509	0.717	**0.816**	0.813	0.799
	Re	0.640	0.544	0.557	0.538	0.632	**0.788**	0.768	0.779
Avg.	F1	0.645	0.535	0.536	0.521	0.645	**0.800**	0.784	0.787
Target: Wearing a Face Mask									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.578	0.692	0.560	0.610	0.640	0.810	**0.840**	**0.840**
	Pr	0.569	0.693	0.551	0.605	0.662	0.803	0.830	**0.835**
	Re	0.580	0.693	0.554	0.603	0.646	0.818	**0.837**	0.819
Avg.	F1	0.567	0.689	0.546	0.599	0.633	0.803	**0.833**	0.825

Table 12: Performance of the baseline models for stance detection on the four targets in the COVID-19-Stance dataset. The performance is reported in terms of accuracy (Acc), precision (Pr), recall (Re) and F1 score (F1). Each baseline was trained and evaluated three times. The results reported are averaged over three runs.

except that there is no softmax layer on top. The discriminators and classifiers were all two layer neural nets with a hidden dimension of 1024. A dropout of 0.15 was used throughout the network. Optimization was performed by AdamW with a learning rate of $3e^{-6}$ for first 7 epochs, and $3e^{-7}$ for the final 3 epochs. The following weights were assigned to this network's loss functions: 0.1 for the domain discriminators, 0.05 for the objective and subjective classifiers, and 0.4 for the source stance classifier. A mini-batch size of 4 was used due to GPU memory limitations.

5.2 Evaluation Metrics

To evaluate the performance of the baseline models on our dataset, we used the following standard metrics: accuracy, (macro average) precision, recall, and F1 score[4]. We report the performance on the test set at the epoch in which the model recorded the highest F1 score on the validation data. We performed 3 independent runs for each model to account for variability, and report average results over the three runs.

[4]Precision, recall and F1 scores for each stance category are also reported in Appendix C

6 Results and Discussion

The results of the experiments are shown in Table 12 for the four targets in the COVID-19-Stance dataset, respectively. Between the two supervised baselines that do not explicitly use the target information, Bi-LSTM and Kim-CNN, the Bi-LSTM gives better results overall, in all metrics, except for the "Wearing a Face Mask" target. When comparing Kim-CNN with GCAE (a CNN-based models that explicitly uses the target), Kim-CNN gives better accuracy and F1 scores for two targets ("Anthony S. Fauci, M.D." and "Stay At Home Orders"), while the GCAE model gives better results for the other two targets ("Keeping Schools Closed" and "Wearing a Face Mask"). Similarly, when comparing the two recurrent models with attention, TAN and ATGRU, TAN performs better on two targets, "Keeping Schools Closed" and "Stay At Home Orders", while ATGRU performs better on "Anthony S. Fauci, M.D." and "Wearing a Face Mask". Surprisingly, these two models, which explicitly use the target information, perform worse than the BiLSTM model overall. Finally, we can see that among the supervised baselines, the BERT model performs significantly better than all the other models, a result that is in agreement with prior works (Ghosh et al., 2019; Miao et al., 2020).

No.	Tweet	Label	NS Prediction	DAN Prediction
1	*@brad_dickson My son teaches in Japan. They wear masks because they are a polite society. School closed asap in Feb. Did remote learning. But as of early June, back to in school learning due to so few cases. Masks work.*	FAVOR	FAVOR	FAVOR
2	*Hell no to your mask mandate*	AGAINST	AGAINST	AGAINST
3	*If, 6 months later, you're still wearing a mask.....you might as well wear one the rest of your life.*	AGAINST	FAVOR	FAVOR
4	*People tweeting from their smart phones about how masks are a form of government control is hilarious to me.*	FAVOR	AGAINST	AGAINST
5	*Thank goodness Trump wasn't there to greet the astronauts after splashdown. I'm sure he would have shown up with no mask! #SplashDown #SpaceX*	FAVOR	FAVOR	AGAINST
6	*Some of ya'll couldn't dissect a frog in high school but you know more than health professionals about the Coronavirus!?!? :man_facepalming_dark_skin_tone: #COVID19*	FAVOR	NONE	FAVOR
7	*Small local grocery store did not have sign requiring mask per state mandate and was pretty busy. Only about half of customers wearing masks. The dairy section looked almost empty. Love it and they will continue to get my business.*	AGAINST	FAVOR	AGAINST
8	*@simondolan @SaltySeaDog7 By not wearing a mask you are giving the children of the COVID generation a chance to go to school, play sports, and have real childhoods*	AGAINST	FAVOR	AGAINST

Table 13: Error Analysis: A comparison of the NS model's predictions with the DAN model's predictions.

When comparing BERT with BERT-NS with BERT-DAN (models that use unlabeled data and SemEval2016 Task 6 data, respectively), we see that BERT performs better than the models that use additional information on the "Stay At Home Orders" target and comparable to the BERT-NS on the "Keeping Schools Closed" target - specifically, the targets with smaller labeled datasets. On the other hand, BERT-DAN performs the best on the "Anthony S. Fauci, M.D." target, and comparable to BERT-NS on the "Wearing a Face Mask" target, i.e., the targets with larger labeled datasets. This result suggests that a larger amount of labeled data is useful for the domain adaptation approach. However, when only a small amount of labeled data is available, BERT is better than the noisy student which may not start with a very good teacher.

Error Analysis. To better understand how two of our best models would perform in the wild, we have included some of their predictions on examples from the Wearing A Face Mask test set, along with the gold-standard label in Table 13. As we can see, both models perform well on examples where the stance is presented explicitly, such as in tweets 1 and 2. However, the models generally struggle with sarcasm and humor as seen in tweets 3, 5, and 6. They also both demonstrate a strong bias towards certain phrases such as "form of government control" which is a common phrase in AGAINST tweets for Wearing A Face Mask. Interestingly, the noisy student model seems to be more likely to incorrectly predict a FAVOR stance when the sentiment of the tweet is positive compared to the DAN model, as seen in tweets 7 and 8.

7 Conclusions and Future Work

In this work, we have constructed a COVID-19-Stance dataset that can be used to further the research on stance detection, especially in the context of COVID-19 pandemic. In addition to the dataset, we have established baselines using several supervised models used in prior works on stance detection, and also two models that can make use of unlabeled data and data from a prior stance detection task, respectively. Our results show the pre-trained COVID-Twitter-BERT model constitutes a strong baseline. When a larger amount of labeled data is available for a target, the BERT-NS and BERT-DAN can help further improve the performance. As part of future work, we plan to study the benefits of the opinion and sentiment data that we annotated towards the stance detection. We also plan to study the usefulness of multi-task learning, where we train models for all our targets concurrently. Other transfer learning approaches that can leverage existing datasets will also be explored.

Acknowledgements

We thank the National Science Foundation and Amazon Web Services for support from grants IIS-1741345, IIS-1802284, IIS-1912887, and IIS-1903963 which supported the research and the computation in this study.

Ethics and Impact Statement

Our dataset does not provide any personally identifiable information as only the tweet IDs and human annotated stance labels will be shared. Thus, our dataset complies with Twitter's information privacy policy. The research enabled by this dataset has the potential to help officials and health organizations understand the public's opinion on various initiatives, estimate the efficacy of their mandates and prevent serious resurgence of the novel coronavirus.

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Charles Courtemanche, Joseph Garuccio, Anh Le, Pinkston J, and Aaron Yelowitz. 2020. Strong social distancing measures in the united states reduced the covid-19 growth rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed covid-19 cases across the united states. *Health Affairs*, 39:10.1377/hlthaff.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

Seth Ebner, Felicity Wang, and Benjamin Van Durme. 2019. Bag-of-words transfer: Non-contextual techniques for multi-task learning. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 40–46, Hong Kong, China. Association for Computational Linguistics.

Marc Evrard, Rémi Uro, Nicolas Hervé, and Béatrice Mazoyer. 2020. French tweet corpus for automatic stance detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6317–6322.

Will Feuer and Nate Rattner. 2020. "Pandemic fatigue" leads to resurgence of coronavirus in europe where cases hit fresh records in france and spain. *CNBC*.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 11696:75–87.

Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in czech news commentaries. In *ITAT*, pages 176–180.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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*.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40, Online. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. 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 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860, Online. Association for Computational Linguistics.

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.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Ece C Mutlu, Toktam Oghaz, Jasser Jasser, Ege Tutunculer, 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. *Data in brief*, 33:106401.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Mike Schuster and Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1868–1873, Minneapolis, Minnesota. Association for Computational Linguistics.

Valeriya Slovikovskaya and Giuseppe Attardi. 2020. Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1211–1218, Marseille, France. European Language Resources Association.

Qingying Sun, Zhongqing Wang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Stance detection via sentiment information and neural network model. *Frontiers of Computer Science*, 13(1):127–138.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. 2016. DeepStance at SemEval-2016 task 6: Detecting stance in tweets using character and word-level CNNs. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 413–419, San Diego, California. Association for Computational Linguistics.

Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in Twitter conversations. *arXiv preprint arXiv:2006.00691*.

Sergey V Vychegzhanin and Evgeny V Kotelnikov. 2019. Stance detection based on ensembles of classifiers. *Programming and Computer Software*, 45(5):228–240.

Penghui Wei, Wenji Mao, and Guandan Chen. 2019. A topic-aware reinforced model for weakly supervised stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7249–7256.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California. Association for Computational Linguistics.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Chang Xu, Cecile Paris, Surya Nepal, Ross Sparks, Chong Long, and Yafang Wang. 2020. Dan: dual-view representation learning for adapting stance classifiers to new domains. In *Proceedings of the 24th European Conference on Artificial Intelligence*, volume 24, pages 2260–2267. European Conference on Artificial Intelligence.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.

Yiwei Zhou and A. I. Cristea. 2017. Connecting targets to tweets: semantic attention-based model for target-specific stance detection. In Athman Bouguettaya, Yunjun Gao, Andrey Klimenko, Lu Chen, Xiangliang Zhang, Fedor Dzerzhinskiy, Weijia Jia, Stanislav V. Klimenko, and Qing Li, editors, *Proceedings of the 18th International Conference on Web Information Systems Engineering, WISE 2017, Puschino, Russia, October 7-11, 2017*, number 10569 in Lecture notes in computer science, pages 18–32. Springer, Cham.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

A Keywords Used for Twitter Crawler

#coronavirus, corona virus, #Coronavid19, #coronavirususa, #coronavirusaustralia, #covid19, covid-19, #covid-19, coronavirus, #coronapocalypse, #quarantinelife, #socialdistancing, SocialDistancing, StayHome, StayAtHome, lockdown, StayHomeSaveLives, Quarantine, socialdistancing, confinement, FlattenTheCurve, StayHomeStaySafe, stayhome, QuarantineLife, 5G, TrumpVirus, StaySafe, Coronavirustruth, WashYourHands, ChineseVirus, TrumpLiedPeopleDied, stayhome, Lockdown, TrumpLiesAboutCoronavirus, ChinaVirus, COVIDIOTS, COVIDIOT, quarantinelife, StaySafeStayHome, hoax, TrumpVirusCoverup, panicbuying, Hydroxychloroquine, TheLockdown, lockdowneffect, toiletpaper, StayAtHomeAndStaySafe, StayTheFHome, SelfIsolation, QuarantineAndChill, stayathome, TrumpPandemic, SocialDistanacing, ChinaLiedPeopleDied, QuaratineLife, lockdownextension, Trumpdemic, TrumpLiedPeopleDied, WorkFromHome, TrumpLiesPeopleDie, QuarentineLife, TrumpLiesAmericansDie, Lockdown21, workingfromhome, TrumpOwnsEveryDeath, TrumpPlague, LockdownExtended, CoronavirusLockdown, TrumpGenocide, SocialDistancingNow, CCPVirus, SocialDistance, ChineseVirus19, ShelterInPlace, StayAtHomeSaveLives, PhysicalDistancing, Resist, Isolation, ChinaCoronaVirus, toiletpapercrisis, lockdownuk, chloroquine, WFH, ChinaLiedAndPeopleDied, LockdownNow, selfisolating, Lockdownextention, CloseTheSchools, Pencedemic, SupportLockdownStaySafe, toiletpaperpanic, schoolclosure, ToiletPaperApocalypse, selfquarantine, masks, handwashing, WearAMask, SafeHands, handsanitizer, LockDown, mask, isolation, flattenthecurve, washyourhands, panicbuyers, panickbuying, Social_Distancing, ChinaMustExplain, Masks4All, WashYourHandsChallenge, BloodOnTrumpsHands, IsolationLife, Hoax, ToiletPaperPanic, toiletpapergate, homeschooling, panicshopping, 5GKILLS, hydroxychloroquine, LockdownHouseParty, trumpvirus, StayHomeSaveLifes, homeoffice, PencePandemic, FamiliesFirst, StayHomeCanada, facemasks, selfisolation, flatteningthecurve, QuaratineAndChill, HerdImmunity, AloneTogether, Hydroxycloroquine, workfromhome, remotework, Masks, FlattenTheCuve, COVIDIDIOT, Socialdistancing, hydroxychloriquine, day8oflockdown, wfh, stayHome, herdimmunity, CoronavirusLockdownUK, TrumpVirus2020, TrumpBurialPits, ShutItDown, 5GCoronavirus, Homeoffice, Resistance, ChineseVirusCorona, chinesevirus, panicbuyinguk, KungFlu, NYCLockdown, facemask, trumpandemic, CoronaHoax, HomeOffice, ChineseCoronavirus, Pandumbic, CoronaLockdown, OPENAMERICANOW, TogetherAtHome, testing, FeverDetectionCamera, WhereAreTheTests, vaccines, Plandemic, Scamdemic, FireFauci, StudentLivesMatter, StayatHome, endthelockdown, ReopenAmerica, lockdown2020, CancelAPExamsPromoteStudents, schoolreopening, HealthOverExams, PromoteStudentsSaveFuture, TestingTestingTesting, schools, lockdownUKnow, SaferAtHome, ContactTracing, FreeThemAll, TrumpCoronavirusTestFailure, TrumpLiedAmericansDied, Handwashing, ChinaLiedPeopleDie, StayAtHomeOrder, OpenAmerica, Vaccine, remoteworking, californialockdown, TestTraceIsolate, EndTheShutdown, WHOLiedPeopleDied, curfew, Curfew, ReOpenAmerica, Testing, TESTVIRUSNOW, socialdistance, plandemic, FakePandemic, stayhomestaysafe, TrumpPandemicFailure, BackToWork, BackToWork, chinavirus, ReopenAmericaNow, MakeChinaPay, TestAndTrace,#MasksOff, MasksOff, SayNoToMasks, #SayNoToMasks, ConstitutionOverCoronavirus, #ConstitutionOverCoronavirus, endthelockdownuk, #endthelockdownuk, #studentban, #StudentBan, SchoolsMustOpeninFall, SchoolReopening, ReopeningSchools, #SchoolsMustOpeninFall, #SchoolReopening, #ReopeningSchools, #Hydroxychloroquine, Hydroxychloroquine

B Questionnaire Used For Amazon Mechanical Turk Workers

Q1: From reading the tweet, which of the options below is most likely to be true about the tweeter's stance or outlook towards stance to prevent the spread of Covid-19:

1. We can infer from the tweet that the tweeter supports the target.
   This could be because of any of reasons shown below:
   - The tweet is explicitly in support for the target.
   - The tweet is in support of something/someone aligned with the target,

from which we can infer that the tweeter supports the target.
- The tweet is against something/someone other than the target, from which we can infer that the tweeter supports the target.
- The tweet is NOT in support of or against anything, but it has some information, from which we can infer that the tweeter supports the target.
- We cannot infer the tweeter's stance toward the target, but the tweet is echoing somebody else's favorable stance towards the target (this could be a news story, quote, retweet, etc).

2. We can infer from the tweet that the tweeter is against the target.
   This could be because of any of the following:
   - The tweet is explicitly against the target.
   - The tweet is against someone/something aligned with the target entity, from which we can infer that the tweeter is against the target.
   - the tweet is in support of someone/something other than the target, from which we can infer that the tweeter is against the target.
   - The tweet is NOT in support of or against anything, but it has some information, from which we can infer that the tweeter is against the target.
   - We cannot infer the tweeter's stance toward the target, but the tweet is echoing somebody else's negative stance towards the target entity (this could be a news story, quote, retweet, etc).

3. We can infer from the tweet that the tweeter has a neutral stance towards the target.
   - The tweet must provide some information that suggests that the tweeter is neutral towards the target - the tweet being neither favorable nor against the target is not sufficient reason for choosing this option. One reason for choosing this option is that the tweeter supports the target entity to some extent, but is also against it to some extent.

4. There is no clue in the tweet to reveal the stance of the tweeter towards the target (support/against/neutral).

Q2: From reading the tweet, which of the options below is most likely to be true about the focus of opinion/sentiment in the tweet:

1. The tweet explicitly expresses opinion/sentiment about the target.
2. The tweet expresses opinion/sentiment about something/someone other than the target.
3. The tweet is not expressing opinion/sentiment about anything.

Q3: What kind of language is the speaker using?

1. The speaker is using positive language, for example, expressions of support, admiration, positive attitude, forgiveness, fostering, success, positive emotional state (happiness, optimism, pride, etc.).
2. The speaker is using negative language, for example, expressions of criticism, judgment, negative attitude, questioning validity/competence, failure, negative emotional state (anger, frustration, sadness, anxiety, etc.).
3. The speaker is using expressions of sarcasm, ridicule, or mockery.
4. The speaker is using positive language in part and negative language in part.
5. The speaker is neither using positive language nor using negative language.

C Comprehensive Results by Class

The average results for stance detection over all three classes, as well as detailed results per class are shown for the four targets in Tables 14, 15, 16, and 17, respectively.

Target: Anthony S. Fauci, M.D.									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.638	0.633	0.588	0.635	0.652	0.817	0.820	0.830
	Pr	0.639	0.685	0.558	0.640	0.661	0.816	0.821	0.833
	Re	0.631	0.612	0.564	0.613	0.634	0.830	0.823	0.839
Average	F1	0.630	0.604	0.547	0.612	0.640	0.818	0.821	0.832
AGAINST	Pr	0.658	0.722	0.588	0.665	0.707	0.859	0.860	0.884
	Re	0.646	0.574	0.585	0.569	0.554	0.841	0.805	0.790
	F1	0.651	0.624	0.584	0.610	0.621	0.850	0.831	0.832
FAVOR	Pr	0.657	0.616	0.623	0.626	0.632	0.860	0.815	0.830
	Re	0.671	0.767	0.715	0.775	0.779	0.739	0.811	0.803
	F1	0.661	0.683	0.655	0.687	0.698	0.795	0.813	0.816
NONE	Pr	0.602	0.716	0.462	0.628	0.643	0.728	0.788	0.785
	Re	0.577	0.494	0.391	0.494	0.571	0.910	0.853	0.923
	F1	0.577	0.506	0.402	0.540	0.601	0.809	0.818	0.848

Table 14: Performance of the baseline models for stance detection on the target "Anthony S. Fauci, M.D.". Average performance over three classes, as well as performance per class is reported in terms of accuracy (Acc), precision (Pr), recall (Re) and F1 score (F1). Each baseline was trained and evaluated three times. The results reported are averaged over three runs.

Target: Keeping Schools Closed									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.627	0.625	0.598	0.590	0.588	0.772	0.780	0.758
	Pr	0.570	0.549	0.545	0.548	0.528	0.765	0.773	0.748
	Re	0.545	0.509	0.532	0.528	0.488	0.761	0.743	0.702
Average	F1	0.548	0.495	0.534	0.527	0.490	0.755	0.753	0.717
AGAINST	Pr	0.372	0.377	0.381	0.370	0.364	0.596	0.660	0.606
	Re	0.238	0.103	0.317	0.310	0.190	0.730	0.651	0.548
	F1	0.287	0.160	0.342	0.321	0.249	0.647	0.652	0.573
FAVOR	Pr	0.674	0.628	0.586	0.616	0.596	0.862	0.869	0.868
	Re	0.594	0.539	0.527	0.545	0.448	0.758	0.709	0.667
	F1	0.629	0.580	0.554	0.572	0.510	0.806	0.779	0.751
NONE	Pr	0.665	0.642	0.667	0.657	0.624	0.836	0.791	0.771
	Re	0.803	0.883	0.751	0.728	0.825	0.796	0.871	0.893
	F1	0.727	0.744	0.706	0.690	0.710	0.813	0.829	0.827

Table 15: Performance of the baseline models for stance detection on the target "Keeping Schools Closed". Average performance over three classes, as well as performance per class is reported in terms of accuracy (Acc), precision (Pr), recall (Re) and F1 score (F1). Each baseline was trained and evaluated three times. The results reported are averaged over three runs.

Target: Stay At Home Orders									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.735	0.703	0.695	0.682	0.738	0.843	0.832	0.833
	Pr	0.679	0.552	0.523	0.509	0.717	0.816	0.813	0.799
	Re	0.640	0.544	0.557	0.538	0.632	0.788	0.768	0.779
Average	F1	0.645	0.535	0.536	0.521	0.645	0.800	0.784	0.787
	Pr	0.700	0.614	0.603	0.613	0.646	0.830	0.784	0.781
AGAINST	Re	0.644	0.615	0.609	0.598	0.701	0.839	0.833	0.839
	F1	0.669	0.614	0.598	0.604	0.671	0.834	0.806	0.809
	Pr	0.785	0.736	0.755	0.721	0.791	0.868	0.869	0.881
FAVOR	Re	0.855	0.881	0.852	0.843	0.849	0.896	0.890	0.881
	F1	0.818	0.802	0.800	0.775	0.817	0.882	0.878	0.881
	Pr	0.554	0.306	0.210	0.194	0.714	0.751	0.786	0.735
NONE	Re	0.420	0.136	0.210	0.173	0.346	0.630	0.580	0.617
	F1	0.447	0.188	0.210	0.183	0.446	0.684	0.667	0.671

Table 16: Performance of the baseline models for stance detection on the target "Stay At Home Orders". Average performance over three classes, as well as performance per class is reported in terms of accuracy (Acc), precision (Pr), recall (Re) and F1 score (F1). Each baseline was trained and evaluated three times. The results reported are averaged over three runs.

Target: Wearing a Face Mask									
		BiLSTM	**Kim-CNN**	**TAN**	**ATGRU**	**GCAE**	**BERT**	**BERT-NS**	**BERT-DAN**
	Acc	0.578	0.692	0.560	0.610	0.640	0.810	0.840	0.840
	Pr	0.569	0.693	0.551	0.605	0.662	0.803	0.830	0.835
	Re	0.580	0.693	0.554	0.603	0.646	0.818	0.837	0.819
Average	F1	0.567	0.689	0.546	0.599	0.633	0.803	0.833	0.825
	Pr	0.613	0.654	0.558	0.589	0.615	0.854	0.859	0.820
AGAINST	Re	0.590	0.735	0.628	0.701	0.765	0.803	0.863	0.936
	F1	0.600	0.691	0.590	0.640	0.675	0.821	0.861	0.874
	Pr	0.436	0.658	0.441	0.530	0.609	0.694	0.766	0.806
FAVOR	Re	0.585	0.699	0.520	0.561	0.667	0.862	0.821	0.707
	F1	0.496	0.674	0.477	0.544	0.619	0.765	0.792	0.753
	Pr	0.658	0.766	0.655	0.696	0.760	0.861	0.864	0.880
NONE	Re	0.564	0.646	0.514	0.547	0.506	0.790	0.827	0.815
	F1	0.606	0.701	0.572	0.613	0.605	0.822	0.845	0.846

Table 17: Performance of the baseline models for stance detection on the target "Wearing a Face Mask". Average performance over three classes, as well as performance per class is reported in terms of accuracy (Acc), precision (Pr), recall (Re) and F1 score (F1). Each baseline was trained and evaluated three times. The results reported are averaged over three runs.