Identifying COVID-19 and Low Income Housing Related Information on Twitter with Semi-Supervised Approaches

Akash Mhatre

Advised by: Dr. Hongmin Li

California State University East Bay, Hayward

Department of Computer Science

`amhatre2@horizon.csueastbay.edu`

1 Introduction

The COVID-19 pandemic has been a global crisis that has affected people of all backgrounds, but some communities have been profoundly impacted by the pandemic. Low-income households [1], in particular, have faced numerous challenges, from the risk of infection due to poor living conditions to job losses due to lockdown measures. These challenges have been compounded by preexisting health issues, limited access to proper healthcare, and difficulty affording basic necessities like food, rent, and utilities.

Moreover, a report from the August 2022 survey has indicated that some groups such as females, transgender individuals, Hispanics, and those without a high school degree, are more likely to experience long COVID, a condition where symptoms persist long after the initial infection has cleared. The rate of self-reported long COVID cases among the above-mentioned groups was 25%-30% higher than those reported globally for all adults [2].

To address these challenges during the pandemic, it is essential to have tools that can monitor the impact of a pandemic on these communities in real-time. One such tool is the 2-1-1 helpline, which provides information about social services such as food assistance, housing support, and mental health resources. Studies of the 2-1-1 helpline data of over 3.5 million requests have indicated that the needs of low-income communities vary greatly across different regions and demographic groups. However, most social needs, apart from unemployment claims, are not systematically monitored by the government in real-time, making it difficult for local government agencies to respond effectively to the needs of their communities [3].

While these new tools and strategies show promise, there are also limitations to survey-based research. Surveys can be expensive and time-consuming to conduct, which makes it difficult for local agencies to monitor the impact of the pandemic on their communities. Furthermore, national agencies may stop tracking the impact of the pandemic at some point, which can limit our ability to prepare for future health crises. Despite these challenges, it is crucial that we continue to develop new tools and strategies for monitoring the impact of the pandemic on low-income communities to ensure that our response is effective.

To address this, Khanal et al. (2021) explored new tools and strategies for monitoring the impacts of the pandemic. Towards this goal, social media has been recognized as a potential real-time source of data for studying and analyzing the impact of the pandemic on low-income communities [4].

1.1 Motivation and Evidence

While survey data is a valuable tool for studying the impact of the COVID-19 pandemic on low-income communities, it suffers from delays that prevent it from providing a real-time indicator of the impact suffered by low-income communities. To address the problem, additional sources of data and information such as social media are being explored as a potential tool. Social media has become a crucial communication tool during the pandemic, as people turned to it to stay connected with family and friends and express their thought and feelings. As a result, social media data can provide a more nuanced understanding of the impact of a pandemic. This can help researchers gain a deeper understanding of the complex and varied impact of the pandemic on low-income communities.

However, collecting and analyzing social media data is not without challenges. One of the biggest challenges is the sheer volume of data available on social media making it difficult to identify relevant data in a time-effective manner, along with noisy and irrelevant or misleading information.

To address these challenges researchers are developing automated tools and algorithms to help identify relevant social media data. Some researchers develop machine learning algorithms and natural language processing techniques that can analyze the impact, identify emerging needs, and come up with strategies to help in real-time. By doing so, these tools can help researchers and policymakers stay informed about the latest developments and respond more effectively to the needs of impacted communities.

1.2 Project Goal

Many studies have focused on analyzing social media data related to COVID-19 pandemic, concerning areas such as sentiment analysis, stance detection, and identifying misinformation. However research focused on analyzing the social media data related to the impact of COVID-19 pandemic has been limited, with some notable exceptions [5, 6, 7, 8, 9, 10, 11].

One such study involved crawling millions of COVID-19-related tweets and manually annotating a small subset of them that best-represented information relevant to low-income households in 16 categories. The annotated tweets were then used to train supervised models based on BERT, a popular language model. While this approach achieved good performance, it only made use of a small subset of the crawled data, leaving a large amount of unlabeled data untapped [4].

To address this, we pursued training a semi-supervised BERT-based model that could leverage both the labeled and unlabeled data to potentially improve the performance of the supervised model for identifying tweets relevant to low-income households, without the need to manually annotate more data. We employed two different strategies for this purpose: Self-Training and Knowledge Distillation [12, 13, 14, 15].

Self-Training [12] involves first training a supervised teacher model on the labeled data, and then using the teacher model to assign "hard" pseudo-labels to the unlabeled data. We then select a subset of the pseudo-labeled data and combine it with the labeled data to train a student model. This process can be iterated for multiple rounds, but the student model may suffer performance loss over time if the pseudo-labeled data is not carefully chosen or if the labels are noisy.

Knowledge Distillation [13, 14, 15], on the other hand, involves using "soft" labels that correspond to the predicted distribution of the unlabeled data. The student model is trained on altering between two objectives: 1) minimizing the cross-entropy loss on the labeled data, and 2) minimizing the cross-entropy loss between the student and teacher-predicted "soft" labels on the unlabeled data.

To summarize, we built semi-supervised BERT models with Self-Training and Knowledge Distillation to better leverage unlabeled data and improve the performance of the supervised BERT model for automatically classifying COVID-19 tweets relevant to low-income households. Our experiment demonstrated that the semi-supervised BERT model and knowledge distillation performed slightly better than the supervised model. These methods have the potential to contribute to the development of more accurate and effective tools for tracking and monitoring the impact of public health crises such as the COVID-19 pandemic, particularly for understanding their effects on vulnerable communities.

1.3 Roadmap Paragraph

"*The following is the outline of the paper. We begin by examining relevant work in Section 2. In particular, we examine the use of tools like surveys, help-line and use of social-media by researchers to understand sentiment and stance of people with respect to social media. We also discussed the different techniques that are used in the study and related work with respect to them. We then provide an overall description of the project in Section 3 as well as the architecture of the of the models that are used in the study along with different training and selection techniques. We also discuss the data collected to train the model along with how the data was labeled. Section 4 we discuss the experiment setup along with the questions we are trying to find answers to through this study. We also discuss the results in Section and examine the accuracy of the models along with how well the models generalizes. Section 5 examines areas for future work. Finally, in Section 6, we summarize the results of the paper and discuss future areas of improvement.*"

2 Related Work

There are many recent works on COVID-19 data analysis tasks, while semi-supervised transformer-based models with Self-Training or Knowledge Distillation have been extensively studied for computer vision and NLP tasks. Given the vast literature on these relevant topics, in what follows, we review papers most closely related to our work.

2.1 COVID-19 Social Media Data Analysis

Many recent studies have focused on collecting COVID-19 social media data, performing content analysis such as stance detection, sentiment analysis, miss information detection and/or automated data analysis using machine learning models. For example, to help emergency services identify risk behaviors, as a means to estimate public mobility with assumptions that mobility reduced risk-averse behavior Senarath et al. (2021) partnered with practitioners to collect and label a dataset of COVID-19 tweets with respect to risk behaviors such as risk-presenting, risk-taking, etc and proposed a machine learning classifier model using lexical and semantic features to classify tweets with respect to behaviors [16]. Imran et al. (2021) presented a TBCOV, a large-scale Twitter dataset comprising more than two billion multilingual COVID-19 tweets collected worldwide over a period of one year [17]. Several deep learning models were used to enrich the data with sentiment labels, named-entities, geolocation and user's gender transformation [17]. Chauhan and Hughes (2021) studied Crisis Named Resources (CNRs) created around COVID-19 on Facebook, Twitter, and Reddit. They analyzed when these resources were created and why, and also how CNR owners attempt to manage content and combat misinformation [6]. Other works used COVID-19 social media data and machine learning for categorization, summarization, sentiment analysis and topic modeling tasks [8, 9, 10, 11, 18, 19]. Most of the survey and studies done for understanding the impact of COVID-19 pandemic on low-income household focus more on social and economic aspect of the impact and having limited resources have their survey and studies stopped after a few months [20]. One of the study [3] showed that there were around 3.5 million 2-1-1 request made but there was not enough data to track the requests and understand the impacts COVID-19 had had on the community making this request. There have been many studies published understanding the sentiment with regards to covid-19 pandemic. Despite such great efforts, limited research has been performed to understand the impacts of COVID-19 pandemic on low-income households through social media data analysis. Most notable, Khanal et al. (2021) [4] crawled millions of COVID-19 tweets, used content analysis to annotate a small subset of them with respect to information relevant to low-income households, and finally built a supervised model based on BERT to automatically identify tweets in 16 categories well-represented in the manually annotated data. The resulting model achieved good performance, without making use of the large amount of unlabeled data readily available. It is of interest to explore semi-supervised approaches that can make use of readily available unlabeled data to potentially improve the results of the supervised models.

2.2 Self-Training

Based on a simple and intuitive idea, Self-Training has been successfully used in many computer vision and NLP tasks Yarowsky 1995, Pise and Kulkarni 2008, Ouali et al. 2020, Zhai et al. 2019 [12, 21, 22, 23]. Self-Training has also been used together with BERT for crisis tweets classification tasks in the context of crisis management. For example, Li et al. (2021) [24] proposed to apply Self-Training with BERT models as base learners to improve performance on domain adaptation tasks where only unlabeled data was available for a target disaster, but labeled data was available for other similar source disasters. They showed that Self-Training could improve the BERT models when the amount of unlabeled data used was relatively large.

2.3 Knowledge Distillation

Hinton et al. (2015) first proposed Knowledge Distillation to compress the knowledge in an ensemble model into a single model for deployment. The conventional approach to Knowledge Distillation involves training a smaller student model to replicate the class probability distributions produced by a larger teacher model [13]. However, recent research has delved into an alternative technique called self-distillation Furlanello et al. (2018); Clark et al. (2019); Zhang and Sabuncu

(2020), where both the teacher and student models possess identical architectures, essentially functioning as a form of semi-supervised learning [25, 26, 27].

Chen et al. (2021) applied both Self-Training and Knowledge Distillation along with two other semi-supervised learning approaches on Natural Language Understanding tasks (specifically, Intent Classification and Name Entity Recognition) [15]. Using a complex data selection procedure and a long short-term memory (LSTM) network as the base learner, they showed that all four semi-supervised approaches reduced the error of the base LSTM model on the tasks considered. Our proposed semi-supervised framework is similar to the framework used by Chen et al. (2021), except that we employed a simpler data selection procedure and replaced the LSTM model with the state-of-the-art BERT model as the base learner. In another closely related work, Zhao and Caragea (2021) used a self-distillation approach with BERT as the base model to learn tag representations for images and subsequently used the tag representations to improve tag-based image privacy prediction. They showed that with only 20% of the annotated data and fine-tuning of the weights associated with the two Knowledge Distillation objectives, the semi-supervised self-distillation approach could achieve performance similar to that of its supervised learning counterpart. Our Knowledge Distillation with BERT approach is very similar to the approach used by Zhao and Caragea (2021) , except that we equally weighted the two objectives of Knowledge Distillation to simplify the training process [14].

3 Description of Project

3.1 BERT models

In this section, we describe in detail the semi-supervised BERT-based approaches used in our study. We also discuss the use of pre-trained BERT based models used for classification of unlabeled data and use of selection process for self-training. We also discuss the dataset used for training and testing the models.

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing (NLP) model developed by Google in 2018. It is based on the Transformer architecture, which is a type of neural network designed to handle sequential data such as natural language. It is able to reads text both left-to-right and right-to-left, allowing it to better understand the context and meaning of words and phrases. It is pre-trained on a large corpus of text, such as Wikipedia and the BookCorpus, and is then fine-tuned on specific NLP tasks, such as question answering, sentiment analysis, and text classification. BERT has achieved state-of-the-art results on many NLP tasks and its pre-trained models have been made available to the public, allowing researchers and developers to fine-tune them for specific applications.

3.2 Semi-supervised BERT models

For a classification task, we assume there exists a labeled dataset $D_l = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ $(i = 1, \ldots, n)$ represent instances and y_i $(i = 1, \ldots, n)$ are their corresponding labels, and also an unlabeled dataset $D_u = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, consisting of instances $\mathbf{x}_j$ $(j = 1, \ldots, m)$ for which the labels are not know. Assuming $m >> n$, a semi-supervised approach will leverage D_u to improve the performance of the model trained only on D_l.

3.2.1 Self-Training (ST)

With ST, we first train a teacher model f_θ^T using D_l. We then use the teacher model f_θ^T to label D_u, and thus each unlabeled instance x_j is assigned a hard (0/1) pseudo-label $\hat{y}_j$. Subsequently, a

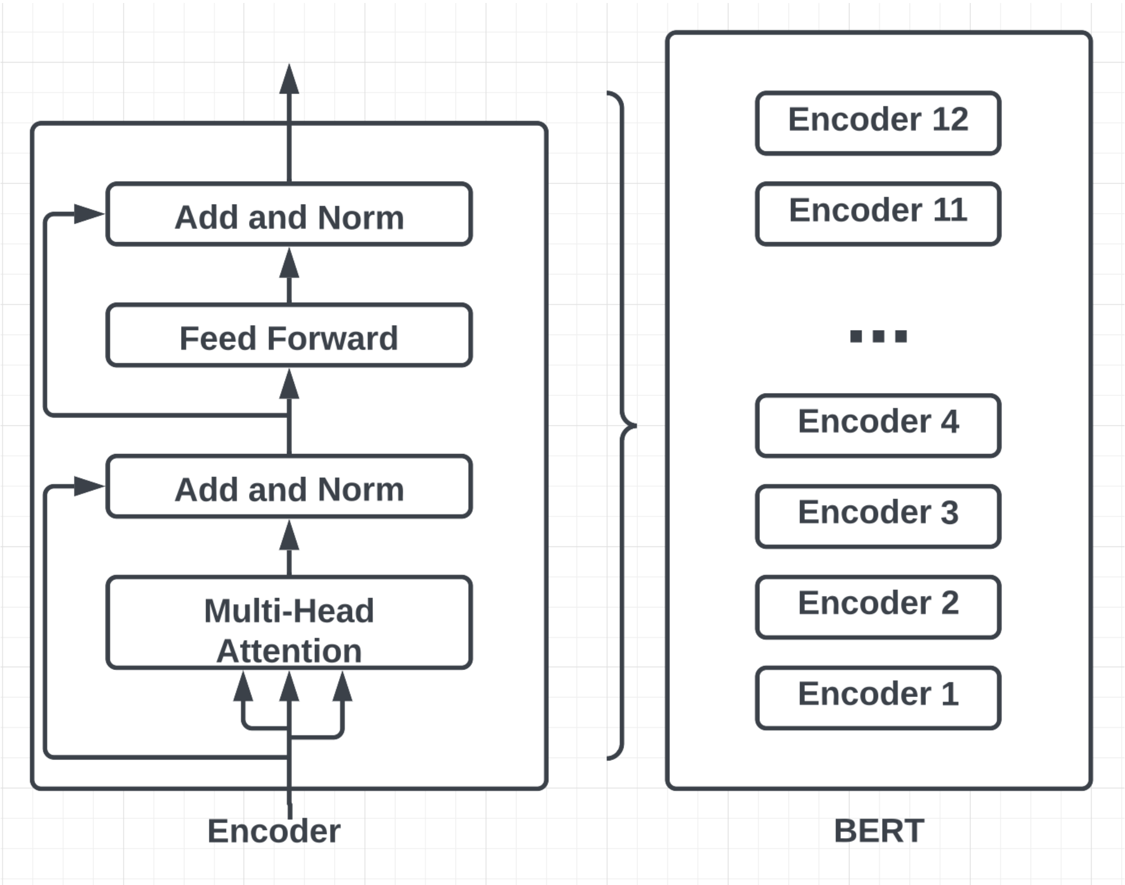

Figure 1: This figure shows the architecture of BERT model model.

new student model f_θ^S is trained on a selected subset of pseudo-labeled instances combined with the originally labeled dataset D_l. This process can be iterated several times, by treating the student model as the "new teacher" and then training a "new student". However, in this paper, we run just one ST iteration (i.e., we trained only one teacher-student pair), and use the student model to make predictions on the test data.

3.2.2 Selection of pseudo-labeled Tweets

To select the best pseudo-labeled tweets for further self-training two methods were used; selection by vote, and selection by average. Tweets with either the highest number of votes or the highest average were selected for further self-training where the pseudo-labeled tweets were combined with labeled tweets for further training.

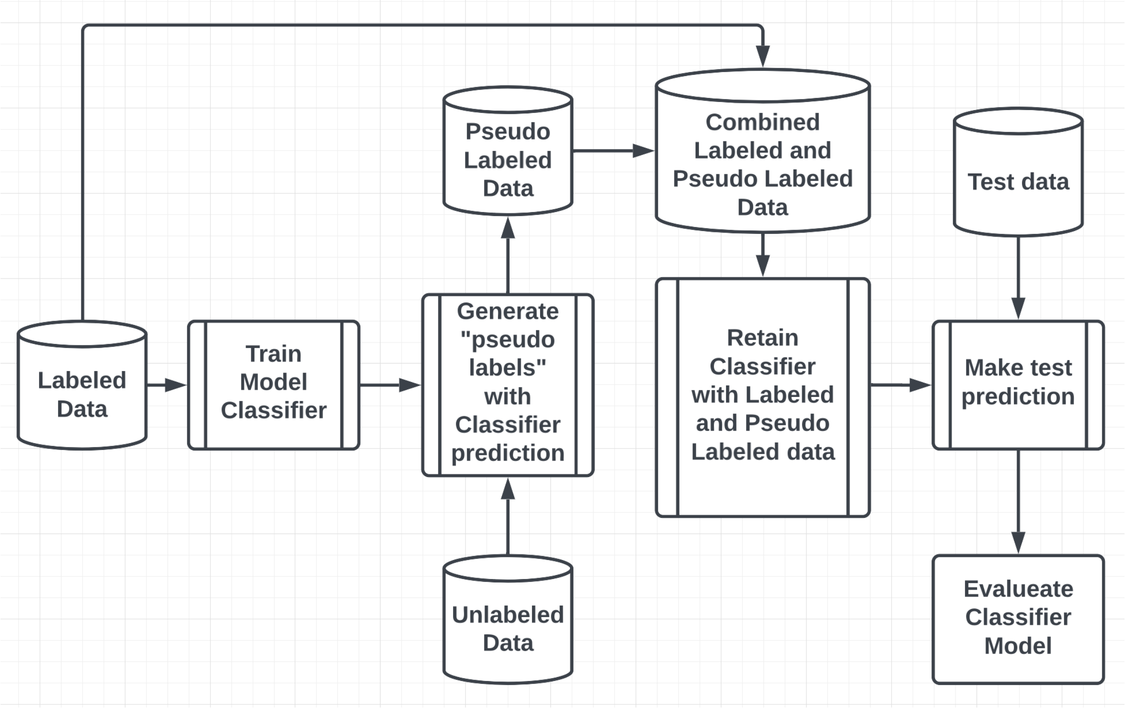
Pseudo Labeled Data
Combined Labeled and Pseudo Labeled Data
Test data
Labeled Data
Train Model Classifier
Generate "pseudo labels" with Classifier prediction
Retain Classifier with Labeled and Pseudo Labeled data
Make test prediction
Unlabeled Data
Evalueate Classifier Model

Figure 2: This figure shows the flow diagram of self-training in a semi-supervised model.

Selection by Vote: In selection by vote, we used the prediction provided by the pre-trained models as a vote. The predicted class receives the vote. The tweet with the most votes for a class is selected for the next step of self-training.

Selection by Average: In selection by average, we will take the average for the prediction by the pre-trained models for each class, and the class with the highest average for a class is selected for the next step of self-training.

3.2.3 Knowledge Distillation (KD)

With KD, we also first train a teacher model f_θ^T on the labeled data D_l only. We then use the teacher model to predict the unlabeled instances $\mathbf{x}_j$ in D_u. However, instead of assigning hard (0/1) labels to the unlabeled instances, the algorithm assigns “pseudo labels” in the form of a probability distribution over classes. For example, for a task with five classes, the pseudo label of an instance might look like $[0.1, 0.1, 0.8, 0.05, 0.05]$. Given these pseudo labels, the student model f_θ^S tries to minimize the difference between its output probability distribution for an instance and the teacher’s predicted probability distribution for that instance. We use the cross-entropy loss to measure the divergence between the two distributions. Thus, the student model is trained using two objectives: minimizing the cross-entropy loss on the original labeled data D_l and minimizing the divergence cross-entropy loss on the unlabeled data D_u with predicted "pseudo labels". Formally, the two objectives (loss functions) are as follows:

$$\mathcal{L}_{sup}(\theta) = \frac{1}{|D_l|} \sum_{(x_i, y_i) \in D_l} l(y_i, p_\theta(y|x_i)) \tag{1}$$

$$\mathcal{L}_{soft}(\theta) = \frac{1}{|D_u|} \sum_{(x_j) \in D_u} l(p_\theta(y|x_j), q(y|x_j)) \tag{2}$$

where l is the cross-entropy loss function, and $p_\theta(y|x_j)$ and $q(y|x_j)$ are the predicted probability distributions of the student and the teacher models, respectively. Finally, the total loss optimized during the training procedure is:

$$\mathcal{L} = \mathcal{L}_{sup}(\theta) + \mathcal{L}_{soft}(\theta) \tag{3}$$

The student model is trained by alternating between minimizing the loss on the labeled data D_l and minimizing the soft-loss on the unlabeled data D_u. More concretely, we train the model with one mini-batch from labeled data D_l, followed by one mini-batch from the soft-labeled data D_u, and continue training until all data is used.

3.2.4 Knowledge Distillation Training Techniques

In the second experiment we train the student model with all the labeled data followed by the soft-labeled data until all the data has been used.

Alternating distillation between Teacher model and Student model: In this method we alternate between training the teacher model using the hard label and then distill the knowledge gained to the

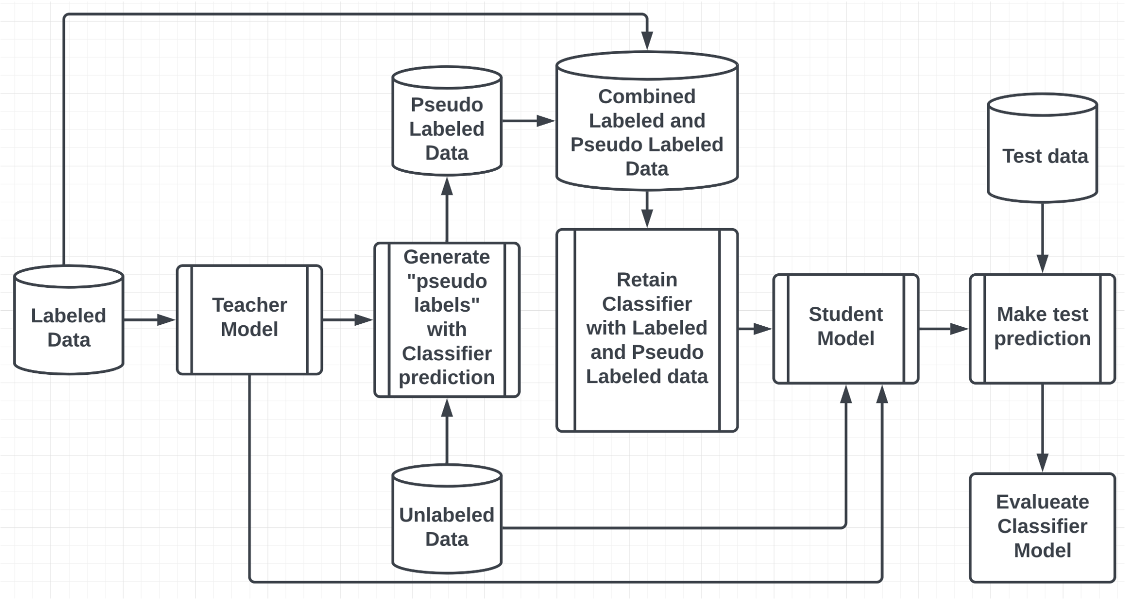
Pseudo Labeled Data
Combined Labeled and Pseudo Labeled Data
Test data
Labeled Data
Teacher Model
Generate "pseudo labels" with Classifier prediction
Retain Classifier with Labeled and Pseudo Labeled data
Student Model
Make test prediction
Unlabeled Data
Evalueate Classifier Model

Figure 3: This figure shows how knowledge distillation in neural network takes place.

student model using hard and soft labels. Both the models are trained simultaneously and its upto the us how the distillation is to be done.

Training Teacher model and distilling it in the Student model: In the following method we first train the teacher model using hard labels followed by distillation the knowledge onto the student model using hard and soft labeled data.

3.3 Pre-Trained BERT Models

We will be making use of some pre-trained models based on BERT (Bidirectional Encoder Representations from Transformers) available on the Hugging Face platform for the pseudo-labeling of unlabeled tweets. It is a type of neural network architecture, to pre-train a language representation model. This pre-training on a large corpus of text allows BERT to learn a rich representation of natural language that can be fine-tuned for a variety of downstream natural language processing (NLP) tasks, such as text classification, question answering, and named entity recognition. To turn BERT-based models into a classification model, we fine-tune it on a labeled classification dataset. Fine-tuning involves taking the pre-trained BERT model and adapting it to a specific task by training it on a smaller dataset. These models are then used for 'soft' labeling and selection. For this, we will be making use of 3 pre-trained models which were specifically trained on COVID-19 data and Twitter's tweets. The pre-trained models would classify the tweets according to the five classes specified, which would then be used for future training of the self-learning model.

3.3.1 BERTweet-COVID

BERTweet-COVID is based on the BERTweet pre-training model. The corpus used to pre-train BERTweet-COVID consists of 850M English Tweets, containing 845M Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related to the COVID-19 pandemic [28].

3.3.2 BERTweet

BERTweet is the first public large-scale language model pre-trained for English Tweets. It is trained based on the RoBERTa pre-training model. The corpus used to pre-train BERTweet consists of 845M English Tweets streamed from 01/2012 to 08/2019 [28].

3.3.3 COVID-Twitter-BERT-v2

COVID-Twitter-BERT-v2 is a BERT-large-uncased model, pre-trained on a corpus of tweets from Twitter about COVID-19. This model is identical to covid-twitter-bert but trained on more data, resulting in higher downstream performance [29].

3.4 Dataset

Khanal et al. [4] used Twitter's Streaming API to crawl tweets that contained keywords pertaining to COVID-19 pandemic such as "#covid-19", "#corona virus", "#covid", etc. They crawled around 170 million tweets posted between March 23rd 2020, to September 25th 2020. Using the keywords that will best represent the tweets from the low-income household community such as "jobs", "CARE Act", "elderly", "low income", "relief" etc., they further segregated the tweets that best represented the tweets from low-income households. Based on the content analysis they did, the tweets were then annotated by human into different categories, for example News and updates related to COVID-19, Stay at home, General public advisory, social inequality and justice etc., 16 classes in total. One thing to notice is that one tweet could be categorized in more than category. But because of some classes have very few tweets, they only built the models for 5 classes: Infected, hospitalized, and deaths (IHD); Social inequality and justice issues (SIJ); Policy responses (personal experiences and opinions toward specific policy) (PLR); Income & economy impacts due to job loss or economics (IEI); Caution and advice to general public (CAG). The number of labeled data is listed in Table 1. The news and update category had the highest number of tweets categorised in it with around 786 tweets, but this is less interesting from low-income households perspective. In addition to the original annotated tweets, after they built the model, they further labeled some additional tweets with human in the loop where they used the model to predicted the tweets labels first and then selected some tweets to ask human annotators to further label these tweets, number of tweets labeled through this human in loop process is shown in Table 1 column of Augmented tweets.

As we will be using semi-supervised learning along with self-training we will use the unlabeled data mentioned in the paper, which is approximately 20, 000. Furthermore, we plan to crawl some more tweets that best represent low-income housing using Twitter's streaming API during the same period to get more unlabeled data to used as input along with the labeled data.

Class	# of Labeled tweets	# of Augmented tweets	Total labeled
Infected, hospitalized, and deaths (IHD)	552	149	701
Social inequality and justice issues (SIJ)	468	137	605
Income & economy impacts (IEI)	359	84	443
Policy responses (PLR)	188	28	216
Caution and advice to general public (CAG)	90	13	103

Table 1: Labeled dataset from [4]

In addition to the labeled dataset, additional 300 labeled tweets were added to the training dataset. These tweets were added to study the impact of adding additional tweets to the original dataset. The additional labeled tweets were manually annotated by 3 coders. The coders understood what each categories represented before labeling additional unlabeled data.

Class	# of Additional Labeled tweets
Infected, hospitalized, and deaths (IHD)	34
Social inequality and justice issues (SIJ)	68
Income & economy impacts (IEI)	86
Policy responses (PLR)	33
Caution and advice to general public (CAG)	79

Table 2: Additional labeled data

3.4.1 Labeling Categories:

- **Infected, hospitalized, and deaths (IHD)**: Tweets that provides information or opinion on the number of people who have been affected by COVID-19, including those who have been exposed, hospitalized, or experienced fatalities, particularly within low-income households and communities. Example: CNN Total Cases Of Coronavirus in 2 Months. (1st March - 1st May) Coronavirus Bar Chart Race USA Surpass 1.1 Million Cases France Surpass 168k Cases UK Surpass 177k Cases #Covid_19 #coronavirus #MayDay2020 #CovidMondial.

- **Social inequality and justice issues (SIJ)**: The tweets that talks about social injustice and inequalities that have been amplified or indirectly caused by COVID-19, including exacerbated poverty, criminal, racial, and other issues. Example: U.S. Asians, harassed over coronavirus, push back on streets, social media.

- **Income & economy impacts (IEI)**: The tweets that cover the effects of COVID-19 on personal finances and on economy, including discussions and complaints about job losses, bankruptcies, and other related topics that may affect low-income individuals. Example: Coronavirus: HSBC puts 35,000 job cuts on hold.

- **Policy responses (PLR)**: The tweets that express either supportive or opposing views towards policies implemented during the COVID-19 pandemic, including comments on specific policies, such as statements regarding noncompliance with mask-wearing, quarantine, or stay-at-home orders, as well as opinions on the implementation of government assistance programs, such as the stimulus package. Example: #StimulusChecks #stimulus #Stimuluscheck #coronavirus #Food House Rent $1800 Utilities $475 Groceries “X” amount and the stimulus check $1200? Intelligent step by Government. Can someone tell them that people need food to live as well? Note:Medical bill not included.

- **Caution and advice to general public (CAG)**: The tweets that offer cautionary advice or recommendation for the general public during the COVID-19 pandemic, including the tweets that report on the issuance or lifting of warnings, as well as guidance and tips to help individuals navigate the current situation. Example: Time to demand accountability from this government for this basura Covid-19 response. Its been 5 months of community quarantine and were worse off than when this started.

4 Experiment setup and Results

4.1 Experiment setup

To compare our semi-supervised BERT model with the supervised model in Khanal et al. (2021), we first reproduced the best supervised model reported in the study. Specifically, we trained a BERTweet-covid model (using a BERT model pre-trained on COVID-19 tweets) that has similar performance as reported in the study Khanal et al. (2021) [4], and used it to establish a baseline for our model. The experiment performed would answer the following questions:

- How does the semi-supervised BERT model with Self-Training compare to the baseline model?
- How does the semi-supervised BERT model with Self-Training and Knowledge Distillation compare to the baseline model?
- How does selection by vote affect the performance of the semi-supervised BERT model?
- How does selection by average affect the performance of the semi-supervised BERT model?
- How does adding more labeled training data affect the performance of the semi-supervised BERT model?

More concretely, we trained and compared the following models:

- **BERTweet-covid (baseline)**: This is the model we built by reproducing the results in Khanal et al. (2021). The model is based on a BERT language model originally pre-trained on tweets Nguyen et al. (2020) [28] and further pre-trained with tweets collected during the COVID-19 pandemic. We trained the model with a batch size of 16, a learning rate of 1e-5, 10 epochs, and callbacks with respect to the best weighted average F1 score on the validation set. We selected the model that had the closest performance to that reported in Khanal et al. (2021). This model serves both as our baseline and as the teacher model for the semi-supervised models [4].

- **BERTweet-covid ST**: This is the semi-supervised model trained with the ST strategy. Specifically, we used the BERTweet-covid as the teacher model, selected the top 500 most confident hard pseudo-labeled tweets for each class (2500 in tweets in total for 5 classes), and combined them with the originally labeled tweets to train the student model. We also explored selecting the top 100/200 most confident tweets, and the model seemed to perform similarly. More careful selection techniques should be explored in future work. We used the same hyper-parameters as for the base BERTweet-covid model, and ran each experiment 5 times with 5 different random seeds. The results for each run and the average results of the 5 runs are reported. We will introduce selection by average and selection by voting to select the best tweets for self-training for training in the next epoch. For selection by average, we will take the average of prediction made by the pre-trained models for each each class and select the best tweets to combine with the training dataset and train the model again. Similarly for selection by vote, we will take the of prediction made by the pre-trained models for each class and convert them into a 0/1 vote (class with highest perdiction value will get 1 vote, remaining classes will get 0) to select the best tweets to combine with the training dataset and train the model again. Selection by vote and selection by average are shown using + vote and + average notation in the result tables.

- **BERTweet-covid ST KD**: This is the semi-supervised model trained with the KD technique by alternating between the two KD objectives described in the Methods section. As for BERTweet-covid ST, we used BERTweet-covid as the teacher model to soft-label the 19,591 unlabeled tweets with predicted probability distributions. We used a batch size of 16 for labeled data, 32 for soft-labeled unlabeled data and trained the model for 5 epochs with callbacks similar to those used for BERTweet-covid ST. We should note that in KD, a temperature T can be applied in the softmax function to smooth the teacher's predicted probability distribution so that the tweets won't be assigned a probability distribution with a value close to 1 for one class and close to 0 for other classes. In our experiments, T was set 1. We also experimented with a value of 10 for T, but the results were not much different.

More extended experiments are needed to see how the T value affects the model. We ran each experiment 5 times and reported results for each run and also average results. We will be performing knowledge distillation in two way and comparing their results with each other. In first, we will train the teacher first before distilling its knowledge to the student. In second case, we will train the teacher and distill the knowledge gained to the student simultaneously. We will first train the teacher for one epoch and distill the knowledge to the student. Then we will train the teacher for two epoch and distill the knowledge to the student and Then we will train the teacher for three epoch and distill the knowledge to the student completing five epochs. Selection by vote and selection by average are shown using + vote and + average notation in the result tables.

4.2 Results

The given data is about the performance of different models based on their precision, recall, F1 score, and weighted F1 score, with results averaged over 5 runs. The data is presented in four different tables, each presenting the results of BERTweet-covid, BERTweet-covid ST, and BERTweet-covid ST-KD models with different selection methods for self-training and different knowledge distillation method. The impact of adding additional labeled tweets to the training data is also shown in the table below.

The results of Table 3 shows that in selection by average BERTweet-covid ST-KD with training teacher followed by student had a higher weighted F1 score of 79.73 which is averaged over 5 runs as compared to the baseline model. The model with training teacher followed by student performed better when compared to BERTweet-covid ST and baseline BERTweet-COVID. Both precision and recall have similar results which is 78.53 and 78.44 respectively. If we use the selection by vote method BERTweet-covid ST-KD using alternating knowledge distillation has higher F1 score and training method with training teacher model followed by student model following close by. The results also shows that in selection by average BERTweet-covid ST-KD with training alternates between teacher and student had a higher weighted F1 score of 79.49 which is averaged over 5 runs

Table 3: Results: Macro Precision, Recall, F1 score and Weighted F1 score of different models. Results are averaged over 5 runs for BERTweet-covid ST and BERTweet-covid ST-KD models with selection of tweets for self-training using selection by average and selection by vote methods.

Model	**Val-F1**	**Precision**	**Recall**	**F1**	**Weighted F1**
BERTweet-covid [4]	83.70	77.23	73.00	75.06	78.51
BERTweet-covid (base/teacher model)	84.29	77.61	74.53	75.48	78.71
BERTweet-covid ST	83.11	74.15	75.97	74.47	77.88
BERTweet-covid KD	83.90	77.83	75.54	76.11	78.90
BERTweet-covid ST + average	77.97	75.13	77.69	75.65	78.07
BERTweet-covid ST-KD training teacher followed by student + average	79.84	78.53	78.44	78.29	79.73
BERTweet-covid ST-KD training alternates between teacher and student + average	79.35	76.8	77.22	76.94	79.28
BERTweet-covid ST +vote	77.32	73.82	76.03	74.53	77.27
BERTweet-covid ST-KD training teacher followed by student + vote	79.19	77.38	77.66	77.37	79.13
BERTweet-covid ST-KD training alternates between teacher and student + vote	79.51	77.8	78.44	77.98	79.49

as compared to the baseline model. The model with training alternates between teacher and student also performed better when compared to BERTweet-covid ST and baseline BERTweet-COVID. From this we can conclude that BERTweet-covid ST-KD performs better than the baseline model with both selection by voting and selection by average using both knowledge distillation training methods.

By analyzing table 4, we can say that adding additional labeled tweet data slight increase in performance in most of the metrics. We can also deduce that selection by voting is performing slightly better than selection by average. We are able to see similar or improved results for recall as well as precision during this experiment. The BERTweet-covid ST-KD model with teacher model training followed by training student model with selection by vote has the best performance amongst all the compared models with 79.38. It is also better than the model trained with no new data. But the same cannot be said for other models as they have provided mixed results where they performed better in few metrics against the similar model with no additional data and sometime perform worst.

Table 4: Results: Macro Precision, Recall, F1 score and Weighted F1 score of different models. Results are averaged over 5 runs for BERTweet-covid ST and BERTweet-covid ST-KD models with selection of tweets for self-training using selection by average and selection by vote methods with additional labeled tweets added to the training set.

Model	**Val-F1**	**Precision**	**Recall**	**F1**	**Weighted F1**
BERTweet-covid [4]	76.83	73.51	75.87	74.07	76.77
BERTweet-covid ST + average	77.48	74.68	76.73	75.13	77.55
BERTweet-covid ST-KD training teacher followed by student + average	79.19	77.95	76.41	76.9	79.04
BERTweet-covid ST-KD training alternates between teacher and student + average	79.19	77.88	77.35	77.44	79.05
BERTweet-covid ST + vote	78.62	75.33	77.45	76.17	78.67
BERTweet-covid ST-KD training teacher followed by student + vote	79.51	77.28	77.58	77.26	79.38
BERTweet-covid ST-KD training alternates between teacher and student + vote	78.86	77.32	76.44	76.55	78.67

The tables show that the BERTweet-covid model with self-training (selection by average) and knowledge distillation (training teacher followed by student) achieved the highest F1 score overall, but the self-training and self-training with knowledge distillation methods were able to slightly improve performance in most cases over the baseline BERTweet-covid model, especially when additional labeled data was used. If the teacher model starts with a noisy mini-batch of tweets, that will make its predication more prone to errors and the teacher will bring in even more noise. It is worth noting that data selection strategies like selection by average and selection by voting helped improve the BERTweet-covid ST-KD model. One could train multiple teachers and filter out noisy data by using a cross-entropy loss of the teachers' predicted probability distributions for the unlabeled data points, or even use a simple threshold-based selection approach.The results suggest that self-training with knowledge distillation can be a useful technique for improving performance on this task. Fine-tuning the hyper-parameters of the models may also help improve the performance.

5 Future Work

Our study has some limitations that we plan to address in future research. We would like to perform extensive experiments with hyper-parameter tuning. Additinally we plan to expand our analysis by collecting tweets related to long COVID and its impact on low-income and other communities. We also believe better data selection techniques could further yield better performance for the semi-supervised models. We would also like to add additional labeled data observe the impact on its performance. Furthermore, we aim to explore the performance of the semi-supervised model when small number of labeled data is provided. This will give us insights into the ability of semi-supervised approaches to reduce annotation costs and address the lack of labeled data. Ultimately, we hope that our future research will contribute to the development of more effective and efficient tools for tracking and monitoring the impact of public health crises, especially in low-income households and other vulnerable communities.

6 Conclusion

The aim of this study was to use semi-supervised BERT models with self-training and knowledge distillation to automatically identify COVID-19 related tweets relevant to low-income housing.

The results of the evaluation showed that the proposed semi-supervised approaches improved the recall of the baseline model by leveraging the unlabeled data. We evaluated our approaches by comparing them to a supervised baseline BERT model proposed by Khanal et al. (2021) [4]. Additionally, the semi-supervised BERT model with self-learning and knowledge distillation demonstrated slight but consistent improvements in all metrics considered. The data selection strategies like selection by average and selection by voting helped improve the BERTweet-covid ST-KD model. This suggests that it can be a useful tool for identifying content relevant to public health crises, particularly those affecting low-income households. It is imperative that we develop such tools to monitor and track the impact of public health crises on low-income household.

Finally, we would like to note that COVID-19 was not the first pandemic that the whole world has faced, nor will it be the last. We hope that our study will contribute to tools that can automatically identify content useful for tracking public health crisis impacts, particularly in relation to supporting the low-income households.

References

[1] K. Parker, R. Minkin, and J. Bennett, “Economic fallout from covid-19 continues to hit lower-income americans the hardest,” PEW RESEARCH CENTER, SEPTEMBER 24, 2020. Retrieved from https://www.pewresearch.org/social-trends/2020/09/24/economic-fallout-from-covid-19-continues-to-hit-lower-income-americans-the-hardest/, 2020.

[2] A. Burns, “Will long covid exacerbate existing disparities in health and employment?” Kaiser Family Foundation, Sep 23, 2022. Retrieved from https://www.kff.org/policy-watch/will-long-covid-exacerbate-existing-disparities-in-health-and-employment/, 2022.

[3] M. Kreuter, R. Garg, I. Javed, B. Golla, J. Wolff, and C. Charles, “"3.5 million social needs requests during covid-19: What can we learn from 2-1-1?",” *Health Affairs Blog*, 2020.

[4] S. Khanal, R. Refati, K. Glandt, D. Caragea, S. Xu, and C.-f. Chen, “Using content analysis and machine learning to identify covid-19 information relevant to low-income households on social media,” in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2021, pp. 1522–1531.

[5] K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea, “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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 1596–1611. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.127

[6] A. Chauhan and A. L. Hughes, “COVID-19 named resources on facebook, twitter, and reddit,” in *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*, A. Adrot, R. Grace, K. A.

Moore, and C. W. Zobel, Eds. ISCRAM Digital Library, 2021, pp. 679–690. [Online]. Available: https://idl.iscram.org/show.php?record=2364

[7] A. Alnuhayt, S. Mazumdar, V. Lanfranchi, and F. Hopfgartner, "Understanding reactions to misinformation - A covid-19 perspective," in *19th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2022, Tarbes, France, May 22-25, 2022*, R. Grace and H. Baharmand, Eds. ISCRAM Digital Library, 2022, pp. 687–700. [Online]. Available: https://idl.iscram.org/show.php?record=2448

[8] Z. Long and R. McCreadie, "Automated crisis content categorization for COVID-19 tweet streams," in *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*, A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel, Eds. ISCRAM Digital Library, 2021, pp. 667–678. [Online]. Available: https://idl.iscram.org/show.php?record=2363

[9] S. Sharma and C. Buntain, "An evaluation of twitter datasets from non-pandemic crises applied to regional COVID-19 contexts," in *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*, A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel, Eds. ISCRAM Digital Library, 2021, pp. 808–815. [Online]. Available: https://idl.iscram.org/show.php?record=2375

[10] A. Evans Jr., Y. Yang, and S. Lee, "Towards predicting COVID-19 trends: Feature engineering on social media responses," in *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*, A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel, Eds. ISCRAM Digital Library, 2021, pp. 792–807. [Online]. Available: https://idl.iscram.org/show.php?record=2374

[11] S. Priya, M. Bhanu, S. K. Dandapat, and J. Chandra, "Mirroring hierarchical attention in adversary for crisis task identification: Covid-19, hurricane irma," in *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM*

2021, Blacksburg, VA, USA, May 2021, A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel, Eds. ISCRAM Digital Library, 2021, pp. 609–620. [Online]. Available: https://idl.iscram.org/show.php?record=2359

[12] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *33rd Annual Meeting of the Association for Computational Linguistics, 26-30 June 1995, MIT, Cambridge, Massachusetts, USA, Proceedings*, H. Uszkoreit, Ed. Morgan Kaufmann Publishers / ACL, 1995, pp. 189–196. [Online]. Available: https://aclanthology.org/P95-1026/

[13] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[14] C. Zhao and C. Caragea, "Knowledge distillation with BERT for image tag-based privacy prediction," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 September, 2021*, G. Angelova, M. Kunilovskaya, R. Mitkov, and I. Nikolova-Koleva, Eds. INCOMA Ltd., 2021, pp. 1616–1625. [Online]. Available: https://aclanthology.org/2021.ranlp-1.181

[15] L. Chen, F. Garcia, V. Kumar, H. Xie, and J. Lu, "Industry scale semi-supervised learning for natural language understanding," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, NAACL-HLT 2021, Online, June 6-11, 2021*, Y. Kim, Y. Li, and O. Rambow, Eds. Association for Computational Linguistics, 2021, pp. 311–318. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-industry.39

[16] Y. Senarath, S. Peterson, H. Purohit, A. L. Hughes, and K. K. Stephens, "Mining risk behaviors from social media for pandemic crisis preparedness and response," in *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation*, 2021.

[17] M. Imran, U. Qazi, and F. Ofli, "TBCOV: two billion multilingual COVID-19 tweets with sentiment, entity, geo, and gender labels," *CoRR*, vol. abs/2110.03664, 2021. [Online]. Available: https://arxiv.org/abs/2110.03664

[18] T. Vijay, A. Chawla, B. Dhanka, and P. Karmakar, "Sentiment analysis on covid-19 twitter data," in *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2020, pp. 1–7.

[19] J.-P. Allem, Q. Li, A. Alasmari, J. Li, T. Ndabu, M. Adly, and A. Adly, "Public perception of the covid-19 pandemic on twitter Sentiment analysis and topic modeling study," *JMIR Public Health Surveill*, 2020.

[20] A. Benton, E. Meade, and A. Vandenberg, "The impact of the first year of the covid-19 pandemic and recession on families with low incomes," *Office of the Assistant Secretary for Planning and Evaluation. US Department of Health and Human Services Institute for Research on Poverty at the University of Wisconsin*, pp. 2021–09, 2021.

[21] N. N. Pise and P. Kulkarni, "A survey of semi-supervised learning methods," in *2008 International Conference on Computational Intelligence and Security, CIS 2008, 13-17 December 2008, Suzhou, China, Volume 2, Workshop Papers*. IEEE Computer Society, 2008, pp. 30–34. [Online]. Available: https://doi.org/10.1109/CIS.2008.204

[22] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *CoRR*, vol. abs/2006.05278, 2020. [Online]. Available: https://arxiv.org/abs/2006.05278

[23] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," 2019. [Online]. Available: https://arxiv.org/abs/1905.03670

[24] H. Li, D. Caragea, and C. Caragea, "Combining self-training with deep learning for disaster tweet classification," in *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*, A. Adrot,

R. Grace, K. A. Moore, and C. W. Zobel, Eds. ISCRAM Digital Library, 2021, pp. 719–730. [Online]. Available: https://idl.iscram.org/show.php?record=2367

[25] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born-again neural networks," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 1602–1611. [Online]. Available: http://proceedings.mlr.press/v80/furlanello18a.html

[26] K. Clark, M. Luong, U. Khandelwal, C. D. Manning, and Q. V. Le, "BAM! born-again multitask networks for natural language understanding," in *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics, 2019, pp. 5931–5937. [Online]. Available: https://doi.org/10.18653/v1/p19-1595

[27] Z. Zhang and M. R. Sabuncu, "Self-distillation as instance-specific label smoothing," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1731592aca5fb4d789c4119c65c10b4b-Abstract.html

[28] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 9–14.

[29] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," 2020.