Identifying COVID-19 Tweets Relevant to Low-Income Households Using Semi-supervised BERT and Zero-shot ChatGPT Models

Hongmin Li[*]
Department of Computer Science
California State University, East Bay
hongmin.li@csueastbay.edu

Doina Caragea[†]
Department of Computer Science,
Kansas State University
dcaragea@ksu.edu

Akash Mhatre
Department of Computer Science,
California State University, East Bay
amhatre2@horizon.csueastbay.edu

Jianye Ge
University of North Texas
Health Science Center
jianye.ge@unthsc.edu

Mingyu Liu
University of North Texas
Health Science Center
muyi.liu@unthsc.edu

ABSTRACT

Understanding the COVID-19 pandemic impacts on low-income households can inform social services about the needs of vulnerable communities. Some recent works have studied such impacts through social media content analysis, and supervised machine learning models have been proposed to automatically classify COVID-19 tweets into different categories, such as income and economy impacts, social inequality and justice issues, etc. In this paper, we propose semi-supervised learning models based on BERT with Self-Training and Knowledge Distillation for identifying COVID-19 tweets relevant to low-income households by leveraging readily available unlabeled data in addition to limited amounts of labeled data. Furthermore, we explore ChatGPT's potential for annotating COVID-19 data and the performance of fine-tuned GPT-3 models. Our semi-supervised BERT model with Knowledge Distillation showed improvements compared to a supervised baseline model, while zero-shot ChatGPT showed good potential as a tool for annotating crisis data. However, our study suggests that the cost of fine-tuning large and expensive GPT-3 models may not be worth for some tasks.

Keywords

COVID Low-income Households, Semi-Supervised Learning, Self-Training, Knowledge Distillation, ChatGPT

INTRODUCTION

The COVID-19 pandemic has profoundly changed many people's lives. While everybody has been affected by the pandemic to some extent, both survey data and research studies have shown that some communities, such as low-income households in the United States, have been disproportionately affected (Parker et al. 2020; Human

[*] corresponding author
[†] corresponding author

Rights Watch 2021; Center on Budget and Policy Priorities 2021; U.S Census Bureau 2023; Kreuter et al. 2020). For example, low-income households and underrepresented communities have been reported to suffer more from long COVID. According to an August 2022 survey, the rates of self-reported long COVID among adults identified as female, transgender, Hispanic or without a high-school degree were 25%-33% higher than those reported globally for all adults. Long COVID was also shown to exacerbate existing disparities in health and employment (Burns 2022; U.S Census Bureau 2023). While the U.S. Census Bureau used surveys to collect data about the needs of vulnerable communities, many critical needs were not systematically tracked by the government during the COVID-19 pandemic, especially not in real-time (Kreuter et al. 2020). To understand COVID-19 impacts on various communities, Kreuter et al. (2020) studied over 3.5 million requests made to 2-1-1, a helpline that provides information about social services, such as help with food, housing, utility bills, mental health, and highlighted that the observed needs varied significantly across different communities. The study emphasized that most social needs, apart from unemployment claims, were not systematically monitored by the government in real-time. Therefore, tools that can help promptly measure and track the impacts of a pandemic on low-income households in real-time will improve the efficiency and effectiveness of the response to low-income households to mitigate the negative impacts, especially the effectiveness of the response from local government agencies that have limited staff and funds. Such tools will also help us to better prepare for future health crises.

Towards this goal, social media has been recognized as a potential real-time source of data for studying the pandemic impacts on low-income households (Khanal et al. 2021). Data from social media platforms can be analyzed and used to come up with strategies for helping the impacted communities, monitoring the effects of policies (Glandt et al. 2021), and tracking long COVID impacts as population's concerns may have shifted with the end of the pandemic. Although social media data is easy and inexpensive to collect in real-time, identifying information relevant to a particular topic is not trivial due to the diversity of topics discussed at the same time on social media, as well as the inherent noisiness of the social media data. In particular, tools that can automatically identify social media data relevant to COVID-19 impacts on low-income households are urgently needed (Khanal et al. 2021).

Many studies on analyzing COVID-19 social media data focused on the use of machine learning models for sentiment analysis, stance detection, misinformation identification, etc. (Chauhan and Hughes 2021; Alnuhayt et al. 2022; Long and McCreadie 2021; Sharma and Buntain 2021; Evans Jr. et al. 2021; Priya et al. 2021; Glandt et al. 2021). However, research focused specifically on analyzing COVID-19 social media data related to low-income households is limited, with some notable exceptions. Among them, Khanal et al. (2021) crawled millions of COVID-19 tweets and used content analysis to manually annotate a small subset of them with respect to information relevant to low-income households. Subsequently, manually annotated tweets in 5 categories were used to train a supervised model based on BERT to automatically identify tweets in those 5 categories. The model developed by Khanal et al. (2021) achieved good performance, although the large amount of unlabeled data crawled was not used.

To leverage readily available unlabeled data, we aim to train semi-supervised BERT-based models which make use of the labeled data from (Khanal et al. 2021) together with unlabeled data to potentially improve the performance of the supervised BERT-based models on the task of identifying tweets relevant to low-income households, without the need to label more data. Furthermore, given the recent ChatGPT[1] buzz and some works showing that ChatGPT could outperform crowd-workers (Gilardi et al. 2023; Kuzman et al. 2023; Huang et al. 2023), we use the most powerful pre-trained ChatGPT model in a zero-shot setting (i.e., no labeled data was provided) to gain insights into its ability to annotate COVID-19 tweets with respect to the 5 categories relevant to low-income households used in this study. Finally, given the cost associated with the use of the GPT API, we explore the utility of paying a small amount of money (in our case, approximately $20) to fine-tune the GPT-3 model based on domain specific data.

More concretely, to train semi-supervised BERT-based models, we use the Self-Training strategy (Yarowsky 1995), and Knowledge Distillation technique (Hinton et al. 2015; Zhao and Caragea 2021; Chen et al. 2021), respectively. Both approaches allow us to incorporate unlabeled data into the training process. With Self-Training, we first train a supervised teacher model, which has performance similar to the performance of the supervised BERT model in (Khanal et al. 2021), and subsequently use the teacher model to assign "hard" pseudo-labels (i.e., 0 or 1 labels) to the unlabeled data. A subset of the data with assigned hard pseudo-labels is selected and combined with the originally labeled data to train a student model. This process can be iterated for some number of iterations, but the student model may suffer performance loss over time if the pseudo-labeled data is not carefully chosen and the labels are noisy. With Knowledge Distillation, instead of hard labels, "soft" labels corresponding to the predicted probability distribution of the unlabeled data are used, and the student model is trained alternating between two objectives: 1) minimizing the cross-entropy loss on the labeled data; 2) minimizing the cross-entropy loss between the student and teacher predicted "soft" labels on the unlabeled data (Chen et al. 2021). Our experiments showed that the student model trained with Knowledge Distillation slightly improves the performance of the original BERT

[1] https://openai.com/blog/chatgpt

teacher model. As opposed to that, with the Self-Training strategy, the student model doesn't always improve the teacher model when we only select the teacher's most confident predictions with hard labels at each iteration.

To experiment with ChatGPT in a zero-shot setting, we used the pre-trained GPT-3.5 model (specifically, gpt-3.5-turbo) and constructed a prompt which asked ChatGPT to classify a given tweet into one of the 5 categories used in our study. We also fine-tuned two GPT-3 models[2] (GPT-3 Ada which is cheaper and faster, and GPT-3 Davinci which is more powerful but also more expensive) for our classification task using the labeled training data. While ChatGPT has been successfully used to perform annotation for other tweet classification tasks (such as identification of hateful tweets or tweet stance detection), the zero-shot results of ChatGPT on our dataset are much better than random guess, but not as good as the results of the fine-tuned BERT model. Similarly, while GPT-3 has shown exceptional performance on many natural language processing (NLP) tasks, the fine-tuned GPT-3 models showed good performance on our data, but not as good as the performance of the fine-tuned BERT model.

To summarize, with the goal of improving the efficiency and effectiveness of the response to low-income households during public health crises such as the COVID-19 pandemic, our main contributions in this paper are the following:

- We built semi-supervised BERT models with Self-Training and Knowledge Distillation to understand their ability to leverage unlabeled data and improve the performance of the supervised BERT model for automatically classifying COVID-19 tweets relevant to low-income households. Our experimental results showed that semi-supervised BERT with Knowledge Distillation performed slightly better than the supervised model. The proposed methods will contribute to build more accurate and efficient tools to track and monitor the impacts of the public health crises like the COVID-19 pandemic.
- We explored the potential of using ChatGPT (with GPT-3.5 model) as an annotation tool for COVID-19 low-income tweets in a zero-shot setting. While ChatGPT has been successfully used for other more general tasks, our results reflected the difficulty of the classification task addressed in this study, suggesting that caution needs to be taken when attempting to use ChatGPT to annotate data for more specific tasks.
- We fine-tuned two GPT-3 models (GPT-3 Ada and GPT-3 Davinci) using our training labeled data to provide the research community with one example of costs versus benefits of the paid GPT models. Our results suggest that the cost of fine-tuning GPT-3 may not justify its benefits, as the fine-tuned models may not surpass the performance of smaller-size open-source fine-tuned language models, such as BERT.

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. Most recently, several notable works have focused on the use of ChatGPT for zero-shot learning (e.g., tweet annotation). Given the vast literature on these relevant topics, in what follows, we review papers most closely related to our work.

COVID-19 Social Media Data Analysis

Many recent studies have focused on collecting COVID-19 social media data, performing content analysis and/or automated data analysis using machine learning models. For example, to help representatives of emergency services identify risk behaviors, as a means to estimate public mobility (under the assumption that mobility is reduced by risk-averse behaviors), Senarath et al. (2021) partnered with practitioners to collect and label a dataset of COVID-19 tweets with respect to risk behaviors (specifically, risk-preventing, risk-taking, or irrelevant) and proposed a machine learning classification framework using both lexical and semantic features to classify tweets with respect to behaviors. Imran et al. (2021) presented TBCOV, a large-scale Twitter dataset comprising more than two billion multilingual COVID-19 tweets collected worldwide over a continuous period of more than one year. Several deep learning models were used to enrich the data with sentiment labels, named-entities, geolocation and user's gender information (Imran et al. 2021). 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 (Chauhan and Hughes 2021). Other works used COVID-19 social media data and machine learning for categorization, summarization, sentiment analysis and topic modeling tasks (Long and McCreadie 2021; Sharma and Buntain 2021; Evans Jr. et al. 2021; Priya et al. 2021; Vijay et al. 2020; Allem et al. 2020). Despite such great efforts, limited research has been performed to understand the impacts of COVID-19 pandemic on low-income households through social

[2] At the time of this writing, OpenAI had released the GPT4 model, in addition to GPT-3.5 model, but GPT-4 and GPT-3.5 were not yet available through the API. GPT-3 model was the most recent model available for fine-tuning through the API.

media data analysis. Most notable, Khanal et al. (2021) first 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 5 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.

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). 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) 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.

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 (Hinton et al. 2015). 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.

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). 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.

ChatGPT

With the great success and fast development of ChatGPT, there are a few works that explore its potential in annotating data (Huang et al. 2023; Gilardi et al. 2023; Kuzman et al. 2023). For example, Huang et al. (2023) examined whether ChatGPT could be used to provide natural language explanations (NLE) for implicit hateful speech detection. Their study showed that ChatGPT could correctly detect 80% of the implicit hateful tweets in the dataset, demonstrating great potential for ChatGPT as a data annotation tool using a simple prompt design. Furthermore, the authors found that the natural language explanations generated by ChatGPT could reinforce human perception, as they tended to be clearer than human-written natural language explanations. Given this, the authors also emphasized the misleading risks posed by wrong answers provided by ChatGPT, and suggested that ChatGPT should to be used with caution. Gilardi et al. (2023) used 2,382 tweets annotated by trained research assistants as gold standard for five tasks (related to discourse around content moderation) to compare ChatGPT's annotations with MTurk crowd-workers' annotations. The tasks included in the study were relevance, stance, topics and two types of frame detection. The results of the study showed that the zero-shot accuracy of ChatGPT exceeded that of the crowd-workers for four out of five tasks. Furthermore, the ChatGPT's intercoder agreement (based on two independent runs) exceeded those crowd-workers and trained annotators for all tasks. Kuzman et al. (2023) examined whether ChatGPT could be used for zero-shot text classification tasks, specifically, genre identification of English and Slovenian tasks. They compared ChatGPT with a multilingual XLM-RoBERTa model fine-tuned on a manually annotated dataset, and found that zero-shot ChatGPT outperformed the fine-tuned XLM-RoBERTa model on both English and Slovenian texts.

METHODS

In this section, we describe in detail the semi-supervised BERT-based approaches used in our study, and also the prompt we designed for experimenting with ChatGPT in a zero-shot setting.

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 $\mathbf{x}_i$ ($i = 1, \dots, n$) represent instances and y_i ($i = 1, \dots, n$) are their corresponding labels, and also an unlabeled dataset $D_u = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, consisting of instances $\mathbf{x}_j$ ($j = 1, \dots, m$) 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.

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 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.

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 "soft labels" in the form of a probability distribution over classes. For example, for a task with five classes, the soft label of an instance might look like $[0.1, 0.1, 0.8, 0.05, 0.05]$. Given these soft 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 "soft 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.

Zero-shot ChatGPT and Fine-tuned GPT-3 models

For zero-shot ChatGPT, we used OpenAI's API for chat completion, where we first sent our prompts to ChatGPT (gpt-3.5-turbo) and then collected the responses as ChatGPT's annotations for the tweets. We started the chat with the following sentence as a system message to the model: "*You are a data expert working on a task to help low-income households impacted by the COVID-19 pandemic.*" Although the official documentation mentions that gpt-3.5-turbo does not pay strong attention to the system message and that the most important information should be put in the prompt, we found that providing the task description helped in our case. Our prompt template is as follow:

You are analyzing tweets and classifying them into one of five categories. The five categories are: 0: 'Infected, hospitalized, and deaths', 1: 'Social inequality and justice issues', 2: 'Income & economy impacts due to job loss or economics', 3: 'Caution and advice to general public', 4: 'Policy responses (personal experience and opinions toward specific policy)'. You only need to give the category number. If you can't tell what it is, say -1.

Tweet: TWEET_TEXT

The classification is:

where TWEET_TEXT was replaced by the actual tweet text. We only asked ChatGPT to categorize the tweets in our test set and compared the zero-shot results with the results of other supervised and semi-supervised models.

To experiment with GPT models fine-tuned on our labeled training data, we chose two paid GPT-3 models, GPT-3 Ada which is the smallest/fastest and least expensive model, and GPT-3 Davinci which is the largest, most capable but also most expensive model. We used the API with default hyper-parameters (e.g., 4 epochs) to save costs.

DATASET

We used the dataset from (Khanal et al. 2021) to evaluate our proposed models on the low-income tweet classification task. Khanal et al. (2021) used Twitter's Streaming API to crawl tweets that contained keywords pertaining to COVID-19 pandemic such as "#covid-19", "#corona virus", "#covid", etc. A dataset consisting of approximately 170 million tweets posted between March 23rd, 2020 and September 25th, 2020 was assembled. Using keywords that best represent the needs of the low-income household community, such as "jobs", "CARE Act", "elderly", "low income", "relief", etc., Khanal et al. (2021) segregated tweets relevant to low-income households and subsequently used content analysis to manually annotate the tweets based on 15 different categories. However, only five categories were included in the dataset used to train and evaluate a BERT model for automated low-income tweet identification (as the other categories had very low representation in the annotated dataset). Specifically, the five categories in the dataset are: 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). Statistics about the dataset are provided in Table 1.

Following the setup in (Khanal et al. 2021), we used 1411 tweets for training, 408 tweets for validation, and the remaining 246 tweets as test dataset. In addition, we used a dataset consisting of 19,591 unlabeled tweets (also provided by (Khanal et al. 2021)) for the semi-supervised approaches.

Category (class)	#Tweets labeled
Infected, hospitalized and deaths (IHD)	701
Social inequality and justice issues (SIJ)	605
Income & economy impacts (IEI)	443
Caution and advice to general public (CAG)	101
Policy responses (PLR)	215
Total	2065

Table 1. Labeled dataset from (Khanal et al. 2021)

EXPERIMENTAL SETUP

To compare our semi-supervised and few-shot models with the supervised models in (Khanal et al. 2021), we first reproduced the best supervised model reported in that work. Specifically, we trained a BERTweet-covid model (using a BERT pre-trained on COVID-19 tweets) that has similar performance to that reported in (Khanal et al. 2021), and used it as the baseline for our models. Our experiments were set up to 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 Knowledge Distillation compare to the baseline model? Also, how does BERT with Knowledge Distillation compare to BERT with Self-Training?
- How does ChapGPT perform on our task in a zero-shot setting? How does the performance of the fine-tuned paid GPT-3 compare with the performance of the baseline BERTweet-covid?

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) and further pre-trained with tweets collected during the COVID-19 pandemic. We trained the model with a batch size of 16, 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.

- **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.
- **BERTweet-covid 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 affect the model. We ran each experiment 5 times and reported results for each run and also average results.
- **Zero-shot ChatGPT**, **GPT-3 Ada and GPT-3 Davinci**. For all three models, we set the temperature hyper-parameter to 0 to make the results more deterministic rather than more random. The GPT-3 models were fine-tuned using default hyper-parameters to keep the fine-tuning costs low.

Following Khanal et al. (2021), the performance of the models was evaluated on the same test set, using the following metrics: (macro) Precision, Recall, and F1 score, and also weighted F1 score.

RESULTS AND DISCUSSION

Table 2 shows the results of the baseline model (BERTweet-covid) by comparison with the results of the semi-supervised models averaged over 5 runs, and the results of the zero-shot ChatGPT and fine-tuned GPT-3 models (one run). More detailed results for the 5 runs of the semi-supervised approaches are shown in Table 3 and Table 4 for the BERTweet-covid KD and BERTweet-covid ST models, respectively.

Table 2. 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 KD models. There is no validation F1 score for zero-shot ChatGPT since it was only run on test set. BERTweet-covid (Khanal et al. 2021) is the baseline model from prior work, and BERTweet-covid (base/teacher model) is the baseline model reproduced in this study.

Model	Val-F1	Precision	Recall	F1	Weighted F1
BERTweet-covid (Khanal et al. 2021)	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**
Zero-shot ChatGPT	-	66.97	61.73	62.83	71.11
GPT-3 Ada Fine-tuned	73.00	71.52	70.71	70.98	76.83
GPT-3 Davinci Fine-tuned	71.70	75.89	73.31	74.39	77.60

As can be seen in Table 2, BERTweet-covid KD performs slightly better than the baseline BERTweet-covid model and also better than the BERTweet-covid ST in all metrics considered. As opposed to that, the results of BERTweet-covid ST are better than those of the baseline only in terms of Recall. This is still a useful result, as Recall is regarded as more important than Precision for the task considered in this study, as it is desirable to not miss relevant low-income information at the cost of including some false positives in the result.

By analyzing the more detailed results of BERTweet-covid KD, shown in Table 3, we can see that BERTweet-covid KD has more significant improvements over the baseline in run 1 and run 5, improving most metrics considered, except for precision in run 5. The results of run 3 are also slightly better than those of the baseline in terms of Recall and F1 but not in terms of Precision and Weighted F1. Run 4 produced relatively similar results to the baseline model, while run 2 resulted in a worse Recall. As described in the Methods section, the KD training process alternates between labeled mini-batches and soft-labeled mini-batches. 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 could help improve our BERTweet-covid KD model, as shown in Chen et al. (2021), who focused on reducing errors for KD-based semi-supervised approaches. For instance, 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. Additionally, fine-tuning the hyper-parameters of the model could potentially improve the model's performance.

Table 3. KD results on each run and also averaged over 5 runs, by comparison with the results of the baseline model

Model	Val-F1	Precision	Recall	F1	Weighted F1
BERTweet-covid (Khanal et al. 2021)	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 KD run 1	84.26	79.75	79.30	79.31	80.43
BERTweet-covid KD run 2	84.26	79.09	69.70	72.42	77.49
BERTweet-covid KD run 3	83.45	76.80	75.97	76.09	78.34
BERTweet-covid KD run 4	83.81	76.11	74.84	75.43	79.20
BERTweet-covid KD run 5	83.73	77.39	77.89	77.29	79.02
BERTweet-covid KD average	83.90	77.83	75.54	76.11	78.90

Table 4. ST results on each run and also averaged over 5 runs, by comparison with the results of the baseline model

Model	Val-F1	Precision	Recall	F1	Weighted F1
BERTweet-covid (Khanal et al. 2021)	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 run 1	82.98	76.76	78.51	77.18	78.55
BERTweet-covid ST run 2	83.14	76.30	77.06	76.61	80.07
BERTweet-covid ST run 3	82.87	74.72	73.88	74.09	77.63
BERTweet-covid ST run 4	83.18	73.36	77.55	74.00	77.37
BERTweet-covid ST run 5	83.39	69.61	72.82	70.48	75.79
BERTweet-covid ST average	83.11	74.15	75.97	74.47	77.88

By analyzing the detailed results of BERTweet-covid ST, shown in Table 4, we can see that only run 2 gave relatively better results than the baseline, while all other runs gave either similar or slightly worse results than the baseline. However, our experiments show that BERTweet-covid ST can increase the Recall by an average of 1.4% and up to 4% in some runs. The addition of 500 hard pseudo-labeled tweets for each class to the original labeled data seems to improve the Recall of the minority classes, but this also leads to a decrease in Precision if the hard pseudo-labels are noisy. Therefore, more careful data selection techniques need to be employed when training semi-supervised models based on the ST strategy. Fine-tuning the hyper-parameters of the models may also help improve the performance.

The results of the zero-shot ChatGPT model, shown in Table 2 show impressive performance considering that no labeled data has been used for training. However, the zero-shot performance of ChatGPT on our task is inferior compared to the performance of the fine-tuned supervised and semi-supervised BERT-based models. This is in contrast to several previous studies that use ChatGPT in a zero-shot setting (Huang et al. 2023; Kuzman et al. 2023; Gilardi et al. 2023) and reported that the ChatGPT performance surpassed the human performance in some cases. We believe the success of ChatGPT in prior works may be due to the nature of the tasks studied, which were either more general language tasks, for example text genre identification (Kuzman et al. 2023) or stance detection (Gilardi et al. 2023), or tasks that the OpenAI team has paid special attention to while training ChatGPT to avoid controversial usage of it, for example hate speech detection (Huang et al. 2023). We found some evidence for this in our trial chats with ChatGPT, when ChatGPT refused to annotate some tweets, citing potential misinformation or hateful speech in the tweet text. Furthermore, our task is more specific, making it more challenging for ChatGPT to be competitive when operating in a zero-shot setting. Nonetheless, we believe that with some carefully chosen examples in the prompts and human in-the-loop, ChatGPT could be used to identify useful crisis information from social media in a zero-shot setting and also to annotate data at a very low cost ($0.002 per tweet using the gpt-3.5-turbo model)[3]. However, caution should be exercised when using the model for specific tasks.

Finally, when analyzing the results of the GPT-3 fine-tuned models in Table 2, we observed that the more expensive model, GPT-3 Davinci gave better results than the cheaper GPT-3 Ada model. However, both models performed worse than the fine-tuned BERT model. BERTweet-covid has the same number of parameters as the original

[3] To be more precise, the cost of ChatGPT with gpt-3.5-turbo is $0.002 per 1000 tokens. For our task, the prompt (including a tweet text) had less than 1000 tokens, so we estimated the cost for annotating each tweet to be $0.002.

BERT base model, which is about 110 million, while GPT-3 Davinci has approximately 175 billion parameters, and GPT-3 Ada model's size is estimated to be 10% of the size of GPT-3 Davinci, which means that GPT-3 Ada has approximately 1.75 billion parameters[4]. Given that the GPT-3 models are much larger than BERTweet-covid, more hyper-parameter tuning may be needed to produce better results. Moreover, much larger training data may be more suitable for fine-tuning GPT-3 given that the model size is much larger. However, it is also important to consider the cost of fine-tuning and running these models for inferences. The cost of fine-tuning and running the GPT-3 Ada model is $0.0004 and $0.0016 per 1000 tokens, respectively, and the cost of fine-tuning and running the GPT-3 Davinci model is $0.03 and $0.12 per 1000 tokens, respectively. For our training data, the cost of one time fine-tuning and using the GPT-3 Ada model for inference was less than $2, while the cost for GPT-3 Davinci model was approximately $7. We experimented with different batch sizes for the GPT-3 Ada model but did not observe a significant change in performance compared to the default batch size. Therefore, based on our GPT-3 exploration, we conclude that the cost of fine-tuning the GPT-3 models many times may not be worth, as they may not outperform the smaller-sized fine-tuned BERT-based models for some tasks, without using significantly larger datasets and incurring larger costs.

CONCLUSIONS

In this paper, we proposed semi-supervised BERT models with ST and KD to automatically identify COVID-19 low-income households related tweets. We also investigated zero-shot ChatGPT and fine-tuned GPT-3 models as alternative solutions for addressing the lack of manually annotated data. We evaluated our approaches by comparing them to a supervised baseline BERT model proposed by Khanal et al. (2021) and found that adding unlabeled data with semi-supervised approaches improved the recall of the baseline model. Additionally, the semi-supervised BERT model with KD showed small improvements compared to the baseline model in all metrics considered. Our study also showed that zero-shot ChatGPT has the potential to be used to annotate crisis data, although caution should be exercised for harder and more specific tasks. Finally, our experiments showed that fine-tuning large and expensive GPT-3 models may not be worth the cost, as smaller fine-tuned BERT models and semi-supervised BERT models gave better results for our task, and similar behaviours may be expected for other similar tasks.

We also discussed the limitations of our experiments, including the fact that no hyper-parameter tuning was performed. Moving forward, we plan to address these limitations by conducting more extensive experiments, such as hyper-parameters tuning for semi-supervised BERT models and designing other prompts for zero-shot ChatGPT to further enhance the results of the analysis in this paper. Furthermore, as the impacts of COVID-19 include the long COVID as well, we plan to collect tweets about long COVID to analyze its impact on low-income households and other vulnerable communities. For the semi-supervised approaches, better data selection techniques can be used to further improve the performance of the models. We will also explore the performance of the semi-supervised models when a smaller number of labeled instances are provided for training to gain better insights into the ability of the semi-supervised approaches to address the lack of labeled data and reduce the annotation costs. Lastly, given that our approaches are based on pre-trained language models, we will also explore few-shot learning with either BERT or ChatGPT as another way to address the challenges posed by lack of labeled data.

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

Allem, J.-P., Li, Q., Alasmari, A., Li, J., Ndabu, T., Adly, M., and Adly, A. (2020). "Public Perception of the COVID-19 Pandemic on Twitter Sentiment Analysis and Topic Modeling Study". In: *JMIR Public Health Surveill.*

Alnuhayt, A., Mazumdar, S., Lanfranchi, V., and Hopfgartner, F. (2022). "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*. Ed. by R. Grace and H. Baharmand. ISCRAM Digital Library, pp. 687–700.

Burns, A. (2022). *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/.

[4] OpenAI has not publicly released the number of parameters number for the GPT-3 Ada model.

Center on Budget and Policy Priorities (2021). *Tracking the COVID-19 Economy's Effects on Food, Housing, and Employment Hardships*. COVID Hardship Watch. Retrieved from https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-economys-effects-on-food-housing-and.

Chauhan, A. and Hughes, A. L. (2021). "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*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 679–690.

Chen, L., Garcia, F., Kumar, V., Xie, H., and Lu, J. (2021). "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*. Ed. by Y. Kim, Y. Li, and O. Rambow. Association for Computational Linguistics, pp. 311–318.

Clark, K., Luong, M., Khandelwal, U., Manning, C. D., and Le, Q. V. (2019). "BAM! Born-Again Multi-Task Networks for Natural Language Understanding". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by A. Korhonen, D. R. Traum, and L. Màrquez. Association for Computational Linguistics, pp. 5931–5937.

Evans Jr., A., Yang, Y., and Lee, S. (2021). "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*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 792–807.

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. (2018). "Born-Again Neural Networks". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by J. G. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1602–1611.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". In: arXiv: 2303.15056 [cs.CL].

Glandt, K., Khanal, S., Li, Y., Caragea, D., and Caragea, C. (2021). "Stance Detection in COVID-19 Tweets". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Association for Computational Linguistics, pp. 1596–1611.

Hinton, G. E., Vinyals, O., and Dean, J. (2015). "Distilling the Knowledge in a Neural Network". In: *CoRR* abs/1503.02531. arXiv: 1503.02531.

Huang, F., Kwak, H., and An, J. (2023). "Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech". In: *CoRR* abs/2302.07736. arXiv: 2302.07736.

Human Rights Watch (2021). *United States: Pandemic Impact on People in Poverty*. Human Rights Watch, March 2, 2021. Retrieved from https://www.hrw.org/news/2021/03/02/united-states-pandemic-impact-people-poverty.

Imran, M., Qazi, U., and Ofli, F. (2021). "TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels". In: *CoRR* abs/2110.03664. arXiv: 2110.03664.

Khanal, S., Refati, R., Glandt, K., Caragea, D., Xu, S., and Chen, C.-f. (2021). "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)*, pp. 1522–1531.

Kreuter, M., Garg, R., Javed, I., Golla, B., Wolff, J., and Charles, C. (2020). ""3.5 Million Social Needs Requests During COVID-19: What Can We Learn From 2-1-1?"". In: *Health Affairs Blog*.

Kuzman, T., Mozetič, I., and Ljubešić, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification*. arXiv: 2303.03953 [cs.CL].

Li, H., Caragea, D., and Caragea, C. (2021). "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*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 719–730.

Long, Z. and McCreadie, R. (2021). “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*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 667–678.

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

Ouali, Y., Hudelot, C., and Tami, M. (2020). “An Overview of Deep Semi-Supervised Learning”. In: *CoRR* abs/2006.05278. arXiv: 2006.05278.

Parker, K., Minkin, R., and Bennett, J. (2020). *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/.

Pise, N. N. and Kulkarni, P. (2008). “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, pp. 30–34.

Priya, S., Bhanu, M., Dandapat, S. K., and Chandra, J. (2021). “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*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 609–620.

Senarath, Y., Peterson, S., Purohit, H., Hughes, A. L., and Stephens, K. K. (2021). “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*.

Sharma, S. and Buntain, C. (2021). “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*. Ed. by A. Adrot, R. Grace, K. A. Moore, and C. W. Zobel. ISCRAM Digital Library, pp. 808–815.

U.S Census Bureau (2023). *Measuring Household Experiences during the Coronavirus Pandemic - Household Pulse Survey*. U.S Census Bureau Household Pulse Survey. Retrieved from https://www.census.gov/data/experimental-data-products/household-pulse-survey.html.

Vijay, T., Chawla, A., Dhanka, B., and Karmakar, P. (2020). “Sentiment Analysis on COVID-19 Twitter Data”. In: *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1–7.

Yarowsky, D. (1995). “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*. Ed. by H. Uszkoreit. Morgan Kaufmann Publishers / ACL, pp. 189–196.

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. (2019). *S4L: Self-Supervised Semi-Supervised Learning*.

Zhang, Z. and Sabuncu, M. R. (2020). “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*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin.

Zhao, C. and Caragea, C. (2021). “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, 1-3September, 2021*. Ed. by G. Angelova, M. Kunilovskaya, R. Mitkov, and I. Nikolova-Koleva. INCOMA Ltd., pp. 1616–1625.