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

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

January 26, 2023

1 Problem Statement

The COVID-19 pandemic has had a profound effect on the lives of many, especially on the low-income households. Many low-income adults have lost their jobs, or had their hours reduced because of stay-at-home orders, spread of COVID-19 virus, pre-existing health care problem etc., making it difficult for them to afford basic necessities such as food, housing, and healthcare. Even with the end of the pandemic, low-income households may suffer more from long COVID, which essentially will impact their social and economic status in the long run. Understanding such impacts during and after pandemic will help the local, state and national agencies to provide necessary supports for these households to recover and also to better prepare for the future public health crisis. There are national agencies' surveys and some research center's limited time surveys that trying to measure such impacts. For example, the United States Census Bureau has conducted The Household Pulse Survey since the pandemic began, where many aspects questions have been asked including the ones for social and economic hardships. However, other types of data is very limited other than this. In this project, we propose to use social media contents as complementary data to study such impacts, which has great potentials especially when surveys are not possible due to resources and funds limit for some local agencies or even national agencies. Social media data is real-time and cheap to collect, but there are challenges to do the analysis as well because it's also very noisy. Currently, there is already research that studies using machine learning models to automatically identify low-income household COVID relevant data. In this project, we propose to use semi-supervised learning approaches to improve such models.

2 Introduction and Motivation

The COVID-19 pandemic has impacted vulnerable communities such as low-income households. These communities are disproportionately affected due to factors such as poverty, lack of access to medical infrastructure, insecure jobs, etc. At the beginning months of the pandemic, according to the research done by PEW Research center until September 2020, since the pandemic started 46% lower-income adults had trouble paying their bills, and 32% said it's been hard for them to make rent or mortgage payments. This percentage is much smaller for middle-income and upper-income adults [1]. Even after receiving government pandemic reliefs aid, according to the United States Census Bureau's Household Pulse Survey from March 2021, among the household with income below $35,000, 47% of the adults reported being behind on rent payments and 25% say they struggle to put food on the table. 32% of the low-income adults said they felt depressed in the previous seven days. These low-income households urgently needed a comprehensive rescue plan [2, 3, 4]. Furthermore, low-income households and underrepresented communities are also suffer more from long COVID. According the August 2022 survey data, rates of self-reported long COVID were 25%-33% higher among adults who were female, transgender, Hispanic, and without a high-school degree than they were among all adults. And this could exacerbate existing disparities in health and employment [5, 4]. Therefore, measuring and tracking such impacts for the long run is important to support the low-income households promptly. However, there could be some limitations to the survey data only based researches. As surveys take a lot of resources and funds to finish and collect, which could make it harder to do for local agencies comparing with national agencies. And even national agencies may stop the tracking at some point. And, the survey data may have some delays to show the true current indicator of the impacts suffered by low-income household. Furthermore, it is not just the adults that are affect by the pandemic, children have also been equally affected by the pandemic due to online teaching and little to no physical contact to the outside world. Understanding and tracking how low-income households children's mental health and also how well they do in school will help us understand the overall impacts of the pandemic on these household as well. Therefore, additional data and tools on COVID low-income household other than the surveys would have great potentials to help study the impacts.

One such tool that can be used to track the impacts is social media. Social media has been the golden mine for many problems since it gets popular. And it's even more true during the pandemic. Because of social distancing and isolation to combat the spread of COVID-19 virus, many have turned to social media to communicate with others and express their feeling and opinion about the situation during the pandemic. Therefore, the data from these social media platforms can be used to analyze the data to come-up strategies to help the impacted communities and to keep monitoring the policies especially economic policies' effects. Although social media is easy and cheap to collect, and it's real-time, it also has a lot of noise in it. So we need good tools to identify COVID low-income household relevant data first and then analyze the filtered data after. According to our knowledge, such tools are still limited [6]. There are many COVID social media researches about sentiment or emotion analysis or topic analysis with supervised machine learning models. But there is only a few studying the specifically on low-income households. Among them, Khanal et al. [6] crawled millions of COVID tweets, then annotated a dataset respect to low-income househould based on their content analysis that have about 15 classes, and finally built a supervised model based on BERT for 5 classes to automatically identify tweets that could belong to these 5 classes. Their model achieved good performance, but a large amount of unlabeled data is not used.

In this project, we propose to use this dataset [6] to build a semi-supervised machine leaning model that will utilize the large amount of unlabeled data to further improve the model. There are many works on semi-supervised learning for many Natural Language Processing (NLP) tasks, we plan to use Self-Training strategy [7], sometimes being referred to as Psuedo-Label (PL), and Knowledge Distillation (KD) [8, 9] as well if time permits. These are the two widely used semi-supervised methods among others. In addition to the unlabeled tweets that's already there, we plan to crawl some additional tweets as well from the same period of this labeled tweets and filter out potential low-income households related tweets with keywords, and use them as unlabeled data in the model. Therefore, our main contributions will be:

- A COVID low-income household tweets unlabeled dataset.
- Reproduce the original paper's result on the labeled data.
- Semi-supervised models with Self-Training (ST) (potentially and Knowledge Distillation (KD)) for COVID low-income household relevant tweets classification, which will be compared to supervised model in Khanal et al. [6].

3 Related Work

There are many works that applied machine learning to the COVID-19 social media data focusing on sentiment analysis and topic modeling. However, not much research is done studying the impact of COVID-19 pandemic on low-income household community through social media content analysis. From the study conducted by M. Kreuter et al. [10] we know that there were around 3.5 million 2-1-1 request made but there was not enough data to track the request and understand the impact covid-19 has had on the community making this request. There have been many studies published understanding the sentiment with regards to covid-19 pandemic. For example T. Vijay et al. [11] worked on the twitter data sentiment analysis by using blob-text module and part of speech text processing to analyze the sentiment of the tweets. Allem et al. [12] studies the public perception of COVID-19 Pandemic over twitter using Latent Duruchlet Allocation technique. Machuca et al. [13] used logistic regression to analyze the sentiment of the tweets related to COVID-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. [14] One study that focused on understanding the impact of covid-19 on low-income household by done by Khanal et al. [6] which focused on content analysis using supervised machine learning models. As they used supervised machine learning model they had to prepare labeled data for training their machine learning model to get results. But in this approach we will be employing semi-supervised machine learning technique to train the model with the help of self training. Using semi-supervised machine learning model will help us reduce the human interaction needed for labeling the data set by combining the small set of labeled data with large set of unlabeled data for training. The model will then use the knowledge from labeled data to label the unlabeled data for training [7, 8, 15, 16, 17].

4 Dataset

Khanal et al. [6] 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., 15 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. In this project, we will the total available tweets as labeled 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 [6]

5 Approach

We first plan to use semi-supervised with Self-Training to build a model, and if time permits we will do semi-supervised learning with Knowledge Distillation as well.

For semi-supervised with self-training, we will train a classifier on the small amount of labeled data, this sometimes being referred to as "teacher" model, and then use the classifier to make predictions on the unlabeled data. The resulting labels for the unlabeled data can be used as "pseudo-labels" which are then given as input to train a new classifier in the subsequent iteration, which is usually referred to as "student" model . This goes on for specified number of iteration or till we achieve a desired amount of iterations. After the classifier is trained with labeled and pseudo-labeled data we test if against a labeled test dataset and evaluate the classifier model [18, 19]. Figure 1 shows the basic process of this approach.

Knowledge Distillation (KD) [8] is another widely used approach int the semi-supervised setting. In this approach, a teacher model will produce a probability distribution over all possible labels for each tweet. The predicted probability distribution is often referred to as “soft label”. The student model is then trained based on two objectives: minimizing the loss on the labeled tweets, and minimizing the cross-entropy loss between the student and teacher predicted “soft label” on the unlabeled data [8]. The objective function could optimize a weighted averaged loss of these two losses, or could also alternating between these two losses with iterations as in [9] for Natural Language Understanding task.

For the base teacher model, BERT based model will be used as in [6]. In the field of research for deep learning for natural language processing transformers have been the most popular architecture. There are many pre-trained transformer models have become available helping in state of the art research in the field of natural language processing. We will be using BERTTweet-covid pretained transformers models, which is based on BERT model but further trained on COVID-19 tweets specifically. We will be comparing the results we get from our semi-supervised model with the supervised model mention in Khanal et al. [6] research paper. The result of the supervised model will be considered as the baseline.

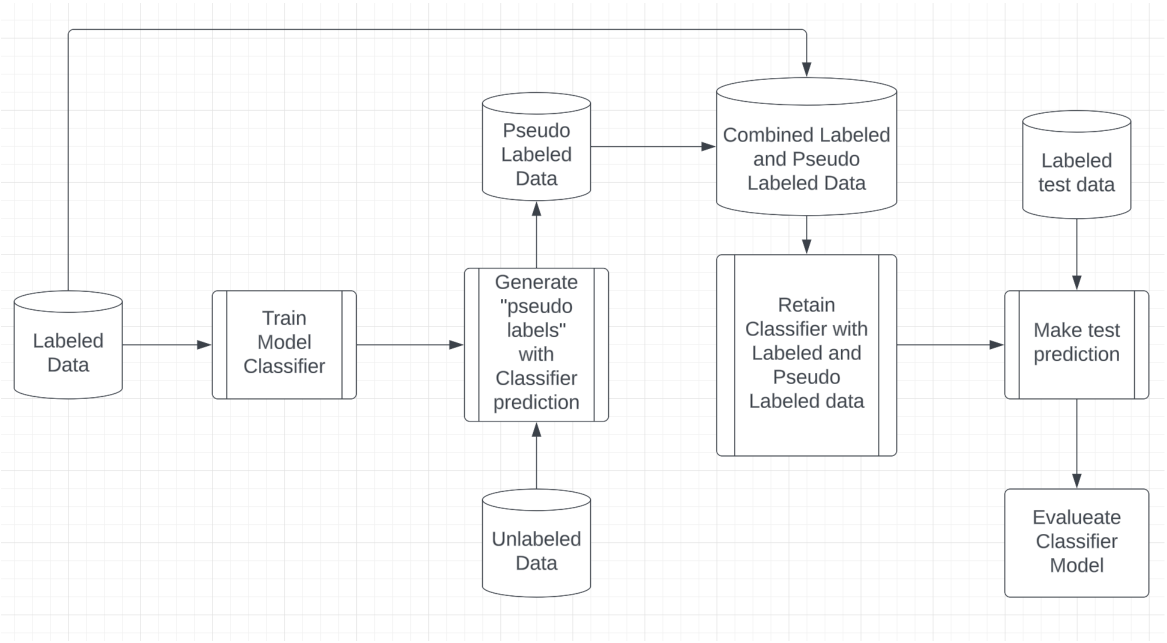
5.1 Schedule

The official class associated with this capstone project should last one semester, which contains roughly 15 weeks of school. Week 1 and Week 2 will be for further literature review. Weeks 3 and 4 will be spent on data collection. Weeks 5 ~7 will be spent on model development and the following 3 weeks (8 ~10) will be spent on hyper parameter tuning . The final fifth of the semester (weeks 13 ~15) will be for documentation, final report, advisor presentation, implementing changes and committee presentation.

Week 1: Literature Review

Week 2: Literature Review

Week 3: Data collection


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

, ,

Figure 1: Training flow diagram for semi-supervised learning model using self-training

Week 4:	Data collection
Week 5:	Model Development
Week 6:	Model Development
Week 7:	Model Development
Week 8:	Hyper parameter tuning
Week 9:	Hyper parameter tuning
Week 10:	Hyper parameter tuning
Week 11:	Comparison between Models
Week 12:	Documentation
Week 13:	Final Report, Advisor Presentation
Week 14:	Feedback Changes
Week 15:	Committee Presentation

6 Project outcomes

The outcome is to create an efficient model that will make use of semi-supervised learning to analysis on the impact COVID-19 pandemic has had on the low-income household community. As the COVID-19 pandemic was not the first crisis faced by the world and it won't be the last crisis that we will face, this tool with the help of transfer learning can be adapted and used in the future to better serve the low-income community to reduce the impact of the public health crisis faced by these and many other community.

References

[1] “Economic fallout from covid-19 continues to hit lower-income americans the hardest,” https://www.pewresearch.org/social-trends/2020/09/24/economic-fallout-from-covid-19-continues-to-hit-lower-income-americans-the-hardest/.

[2] “United states: Pandemic impact on people in poverty,” https://www.hrw.org/news/2021/03/02/united-states-pandemic-impact-people-poverty, 3/2/2021.

[3] “Tracking the covid-19 economy’s effects on food, housing, and employment hardships,” https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-economys-effects-on-food-housing-and.

[4] “Measuring household experiences during the coronavirus pandemic - household pulse survey,” https://www.census.gov/data/experimental-data-products/household-pulse-survey.html.

[5] “Will long covid exacerbate existing disparities in health and employment?” https://www.kff.org/policy-watch/will-long-covid-exacerbate-existing-disparities-in-health-and-employment/.

[6] 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.

[7] 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/

[8] 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

[9] 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

[10] 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.

[11] 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.

[12] A. A. J. L. T. N. M. A. Jon-Patrick Allem, Qiudan Li and A. Adly, “Public perception of the covid-19 pandemic on twitter Sentiment analysis and topic modeling study,” *JMIR Public Health Surveill*, 2020.

[13] C. R. Machuca, C. Gallardo, and R. M. Toasa, “Twitter sentiment analysis on coronavirus: Machine learning approach,” *Journal of Physics: Conference Series*, vol. 1828, no. 1, p. 012104, feb 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1828/1/012104

[14] “The impact of the first year of the covid-19 pandemic and recession on families with low incomes,” https://www.pewresearch.org/social-trends/2020/09/24/economic-fallout-from-covid-19-continues-to-hit-lower-income-americans-the-hardest/.

[15] Z.-H. Zhou, *Semi-Supervised Learning*. Singapore: Springer Singapore, 2021, pp. 315–341. [Online]. Available: https://doi.org/10.1007/978-981-15-1967-3_13

[16] Y. Ouali, C. Hudelot, and M. Tami, “An overview of deep semi-supervised learning,” 2020. [Online]. Available: https://arxiv.org/abs/2006.05278

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

[18] N. N. Pise and P. Kulkarni, “A survey of semi-supervised learning methods,” in *2008 International Conference on Computational Intelligence and Security*, vol. 2, 2008, pp. 30–34.

[19] O. T. Nartey, G. Yang, J. Wu, and S. K. Asare, “Semi-supervised learning for fine-grained classification with self-training,” *IEEE Access*, vol. 8, pp. 2109–2121, 2020.