What are We Depressed about When We Talk about COVID-19: Mental Health Analysis on Tweets Using Natural Language Processing

Irene Li[1], Yixin Li[1], Tianxiao Li[1], Sergio Alvarez-Napagao[2], Dario Garcia-Gasulla[2], and Toyotaro Suzumura[2]

[1]Yale University, USA
[2]Barcelona Supercomputing Center (BSC), Spain

Abstract. The outbreak of coronavirus disease 2019 (COVID-19) recently has affected human life to a great extent. Besides direct physical and economic threats, the pandemic also indirectly impact people's mental health conditions, which can be overwhelming but difficult to measure. The problem may come from various reasons such as unemployment status, stay-at-home policy, fear for the virus, and so forth. In this work, we focus on applying natural language processing (NLP) techniques to analyze tweets in terms of mental health. We trained deep models that classify each tweet into the following emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. We build the EmoCT (Emotion-Covid19-Tweet) dataset for the training purpose by manually labeling 1,000 English tweets. Furthermore, we propose an approach to find out the reasons that are causing sadness and fear, and study the emotion trend in both keyword and topic level.

Keywords: Deep Learning · Mental Health · Natural Language Processing · Topic Modeling

1 Introduction

Mental health is becoming a common issue. According to World Health Organization (WHO), one in four people in the world will be affected by mental or neurological disorders at some point in their lives[1]. A large emergency, such as the coronavirus disease 2019 (COVID-19), would especially sharply increase people's mental health problems, not only from the emergency itself, but also from the subsequent social outcomes such as unemployment, shortage of resources and financial crisis. Almost all people affected by emergencies will experience psychological distress, which for most people will improve over time[2]. In order

[1] https://www.who.int/whr/2001/media_centre/press_release/en/
[2] https://www.who.int/news-room/fact-sheets/detail/mental-health-in-emergencies

to help the society get prepared in response to surging mental problems during and after COVID-19 emergency, we need to understand people's general mental status as a first step.

Language, as a direct tool for people to convey their feelings and emotions, can be very helpful in the estimation of mental health conditions. Due to the recent impact of COVID-19, a large number of people move their works online, making some users are more active than usual on their social media accounts. Previous works have been conducted to utilize natural language processing (NLP) methods to process internet-based text data such as posts, tweets, and text messages on mental health problems [2,4,9,6].

There are mainly two challenges in working with tweets using NLP methods. The first challenge is the large number of new posts online but restricted availability of APIs. There may be up to 90 or even 100 million tweets per day [4], so most of research is conducted on random samples [14,10,12]. We are interested in a million-level of tweets and also in a larger time span. Another challenge is the lack of labeled dataset for COVID-19. Though there exist labeled Twitter dataset for sentiment and emotions [7,10,8], due to the domain discrepancy, we still wish to have a manually-labeled dataset for training to have a better performed model. The work by [9] applies principal component analysis (PCA) to predict emotions. The method by [1] is to use k-Nearest Neighbors and Naive Bayes classifier to do classification on tweets. Very recently, many types of contextualized word embeddings are proposed and substantially improved the performance on many NLP tasks. A new language representation model, BERT [5], was proposed and obtains competitive results on up to 11 NLP tasks like classification. In this work, we apply a pre-trained BERT and fine-tune on our labeled data, providing in-depth analysis of mental health.

Our contributions are three-fold: we build the **EmoCT** (**Emo**tion-**C**OVID19-**T**weet) dataset for classifying COVID-19-related tweets into eight emotions; then, we propose two models to do both single-label and multi-label classification respectively based on a multilingual BERT model, which are capable to predict on up to 104 languages and achieving promising results on English tweets; further analysis on case studies provide clues to understand why and how the public may feel fear and sad about COVID-19, and we also provide the study of emotion trend in both keyword and topic level.

2 Dataset

We applied Twitter API[3] to conduct a crawler with a list of keywords:coronavirus, covid19, covid, COVID-19, covid_19, confinamiento, flu, virus, hantavirus, fever, cough, social distance, lockdown, pandemic, epidemic, conlabelious, infection, stayhome, corona, épidémie, epidemie, epidemia, 新冠肺炎, 新型冠病毒, 疫情, 新冠病毒, 感染, 新型コロナウイルス, コロナ. Each day, we are able to crawl 3 million tweets in free text format from different languages. Due to the high

[3] `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json`

capacity, we look at the tweets from March 24 to 26, 2020 to get language and geolocation statistics. Among these tweets, 8,148,202 tweets have the language information (`lang` field of the `Tweet` Object in Tweet API), and 76,460 tweets have the geographic information (`country_code` value from the `place` field if not `none`). We show the geolocation distributions in Fig 1. Besides, Fig 2 shows the language distribution of 8,148,202 tweets from 24 to 26 March, 2020 from the `lang` field of the tweet object.

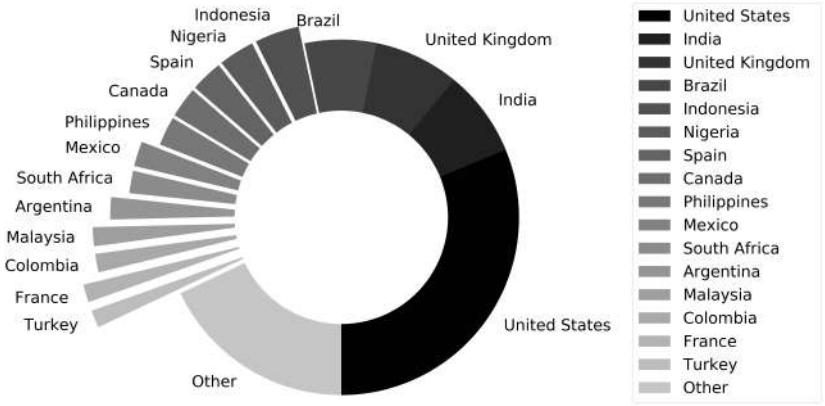

Fig. 1: Geolocation distribution on 76,460 tweets.

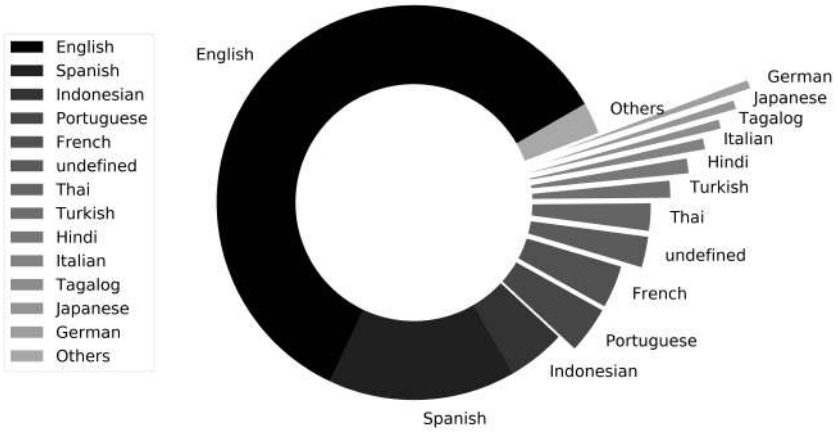

Fig. 2: Language distribution on 8,148,202 tweets.

To investigate the problem of mental health, we come up with the task of emotion classification on COVID-19-related tweets. We built **EmoCT** (**Emo**tion-**C**ovid19-**T**weet) dataset. We randomly annotated 1,000 English tweets selected from our crawled data. Following the work of EmoLex [11], we classify each tweet into the following emotions: anger, anticipation, disgust, fear, joy, sadness,

surprise and trust. Each tweet is labeled as one, two or three emotion labels. For each emotion, we made sure that the primary label appeared in 125 tweets, and there is no number control in the secondary and tertiary label. We then split into 100/25 for each emotion as the training/testing set. We release two versions of the dataset: single-labeled version where only the primary label is kept for each example, and multi-labeled version where all the labels are kept. In this way, both single-label classification and multi-label classification can be conducted. We release the dataset to the public [4].

3 Classification

Single-label Classification We first attempt to do a single-label classification task based on the single-labeled version of EmoCT. We apply a pre-trained multilingual version BERT model[5]. We take the output of the `[CLS]` token and add a fully-connected layer, which is fine-tuned using the labeled training examples (BERT). We set the learning rate to be 10^{-5} and number of epochs to be 20. Besides, we also fine-tune with the MLM (masked language model) on 1,181,342 unlabeled tweets randomly selected from our crawled data, and then trained on EmoCT (BERT(ft)). Tab 1 shows the performance of the two models. As we can see, both models have competitive results on accuracy and F1, and BERT(ft) performs slightly better than BERT, so we take this model as our main model for analysis in later sections.

Method	Accuracy	F1
BERT	0.9549	0.9545
BERT(ft)	0.9562	0.9558

Table 1: Single-label Classification Results on EmoCT single-labeled version.

Method	Average precision	Coverage error	Ranking loss
BERT	0.6415	3.2261	0.2325
BERT(ft)	0.6467	3.1256	0.2159

Table 2: Multi-label Classification Results on EmoCT multi-labeled version.

Multi-label Classification We also perform multi-label classification on the multi-labeled version of EmoCT. In this setting, each tweet has up to three

[4] `https://github.com/IreneZihuiLi/EmoCT`

[5] https://github.com/huggingface/transformers: *bert-base-multilingual-cased* model

	BERT	BERT(ft)
Anger	0.7473	0.7397
Anticipation	0.6173	0.6897
Disgust	0.8222	0.8364
Fear	0.7010	0.7344
Joy	0.8380	0.8430
Sadness	0.7394	0.6809
Surprise	0.8620	0.8676
Trust	0.7919	0.8228
Micro-Avg.	0.7778	0.7891

Table 3: AUROC for each label of multi-label Classification on EmoCT multi-labeled version, as well as the micro average over all classes.

labels out of eight, and we assume the labels are independent. We build a single-layer classifier with the activation function to be Sigmoid, which receives BERT output and predicts the possibility of containing each of the eight labels (BERT). The model uses binary cross-entropy loss and is trained for 10 epochs with learning rate 10^{-5}. Similarly, we also compare with a fine-tuned version as did in the previous model (BERT(ft)). For evaluation, we use example-based evaluation metrics mentioned in the work of [15] in Tab 2. We could see that the two models achieve relatively low scores, probably due to the small-scale training data. We leave the improvement of this model as future work. In Tab 3, we show the area-under-curve (AUC) of the response operating characteristic (ROC) curve for each class and their micro average. It can be noticed that both models are not performing so well by looking at the average score, and they are not very confident on certain classes like *anticipation*, and we leave it as future work.

4 Correlation

Due to the outbreak of coronavirus emergency, the two emotions *sad* and *fear* are more related to severe negative sentiments like depressed. To understand why the public may feel fear and sadness, we then attempt to analyze words and phrases that have a high correlation with both emotions. We apply our BERT(ft) model from the single-label classification task to predict the emotion label on randomly-picked 1 million English tweets data on April 7, 2020. Note that we keep only the tweets labeled as fear and sadness.

4.1 Attention Weight

When predicting the emotion label for each tweet, we take the last attention layer of the model and collect the top 3 tokens which have the maximum attention weights. Finally we rank the tokens by frequency and plot the wordcloud[6] of

[6] Visualization tool: https://wordart.com/. Invalid for a few languages.

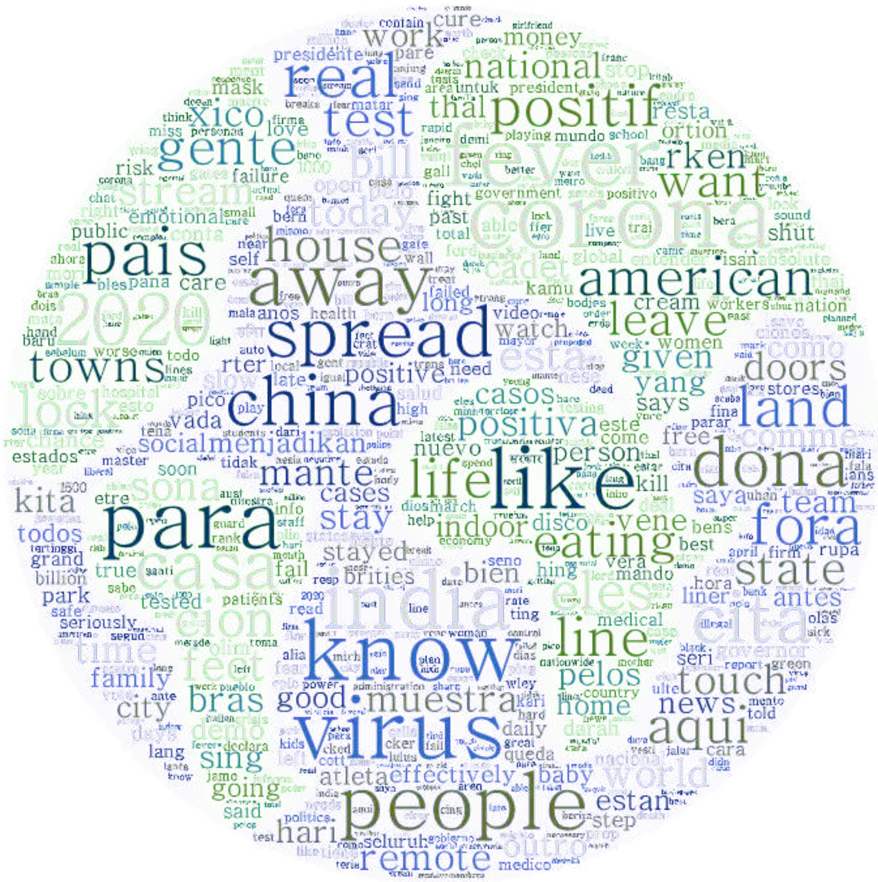

Fig. 3: Wordcloud from attention.

the top 500 tokens after filtering some stopwords in Fig 3. A drawback of this method is that the tokens are split, so we can see some keywords that may not be meaningful without contexts, for example: *like*, *know* and *2020*. However, we can get some reasonable keywords: *fever*, *corona*, *spread*, *virus* and so on. Such words appear with a high frequency in the tweets labeled as fear and sadness, which may explain what and why people are feeling fear or sad. Note that this method can handle multiple language input as the pre-trained BERT model supports 104 different languages, though training was conducted on an English corpus.

4.2 POS tagging

Intuitively, we assume that nouns are more meaningful in a tweet, making it possible and easier to understand the reasons why it is labeled as fear or sad-

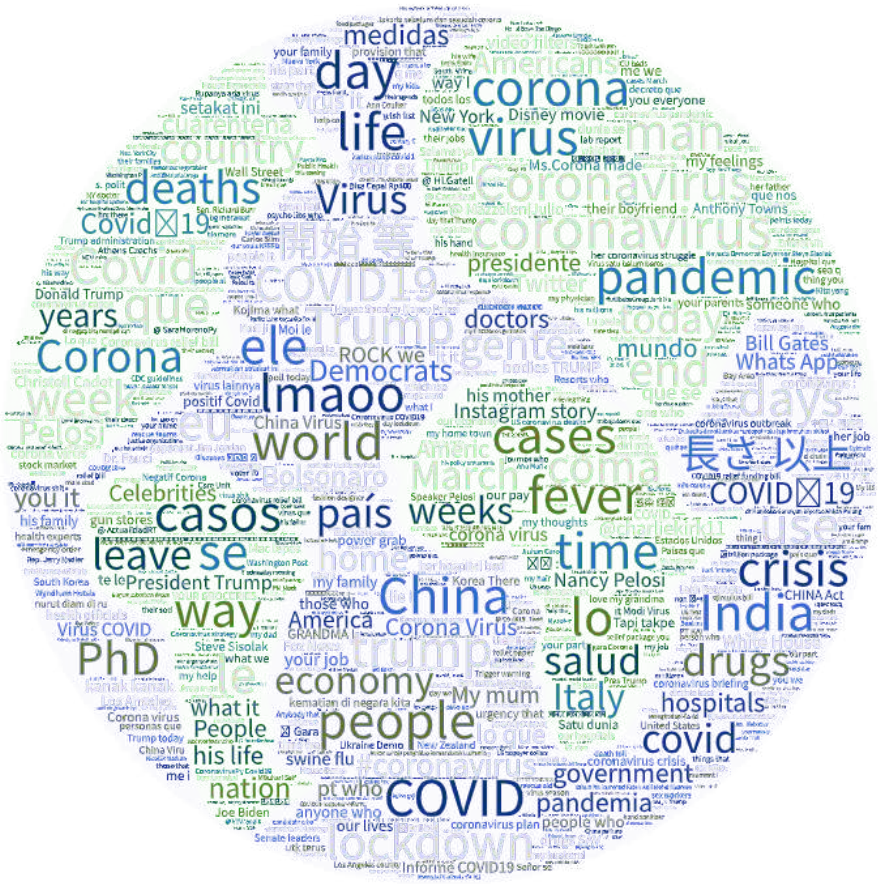

Fig. 4: Wordcloud from POS labelging.

ness. As a comparison, we look at the Part-of-Speech (POS) tag of each token in the tweets and keep the nouns and noun phrases only. We apply the Stanza Python library to do POS tagging [13] and we include supporting to six languages including English, Spanish, Portuguese, Japanese, German and Chinese. Similarly, we plot the top 500 keywords and phrases based on frequency in Fig 4. There are some informative keywords and phrases captured: *pandemic*, *China*, *economy*, 開始 (means *starting* in English), *President Trump*, *White House* and so on. While working on the analysis, we saw other meaningful phrases such as *gun stores*, *school closings*, and *health conditions* which has a lower frequency and may not be visible.

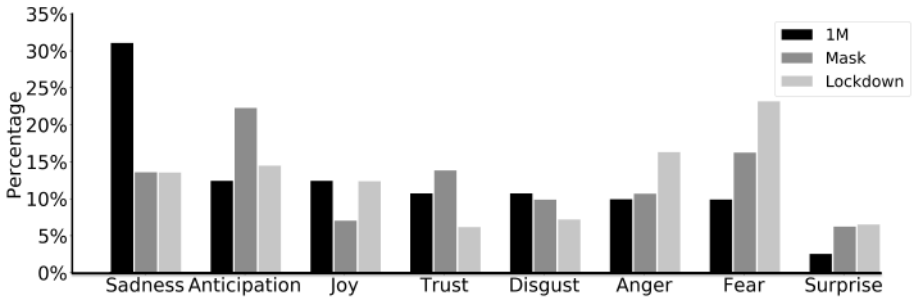

Fig. 5: Comparison of emotion distribution.

5 Emotion Trend Analysis

The emotion trend among different keywords or topics is also very important, as it potentially may show the public attitude change within a period of time. Similarly, we provide analysis only on the English tweets.

5.1 Emotion trend with keywords

We still choose the single-label classification BERT(ft) model to do prediction. We provide a case study on two keywords: *mask* and *lockdown*. We first pick 1 million tweets randomly from the data of March 29th, 2020. By filtering on the keywords, we found 8,071 tweets that contain the word *mask*, and 31,146 tweets that contain the word *lockdown*. Fig 5 shows the comparison of emotion distribution among 1 million samples (1M), tweets with *mask*, and tweets with *lockdown*. In the 1M group, most tweets are classified into negative classes like fear, anger and sadness. But when people are talking about masks, more tweets are classified into anticipation and trust, which is sometimes more neutral and positive. For the tweets talking about lockdown, there is no significant difference with that of 1M.

To further analyze the trends, we select the data of two weeks (March 25, 2020-April 7, 2020), and apply the same model to predict the emotion labels on all the tweets we crawled (around 3 million each day) that contains the two mentioned keywords respectively. There is no significant change for the emotion distribution in all the data. However, we found the dominating emotions and variations of the change are closely related to the topic. In Fig 6 and 7, we illustrate the emotion trend for each single day of the selected keywords. The high variation (plot in solid lines in the figures) showed up in *sadness*, *anger* and *anticipation* for the tweets that contain the word *mask* in Fig 6, and *disgust*, *sadness* for the tweets that contain the word *lockdown* in Figure 7. Especially, for the *lockdown* tweets, the percentage of *disgust* emotion had a significant increase on March 27 and dropped on the next two days, as marked with the black asterisks. To further investigate, we looked at the news in March 27, which included U.S. as the first country to report 100,000 confirmed coronavirus cases,

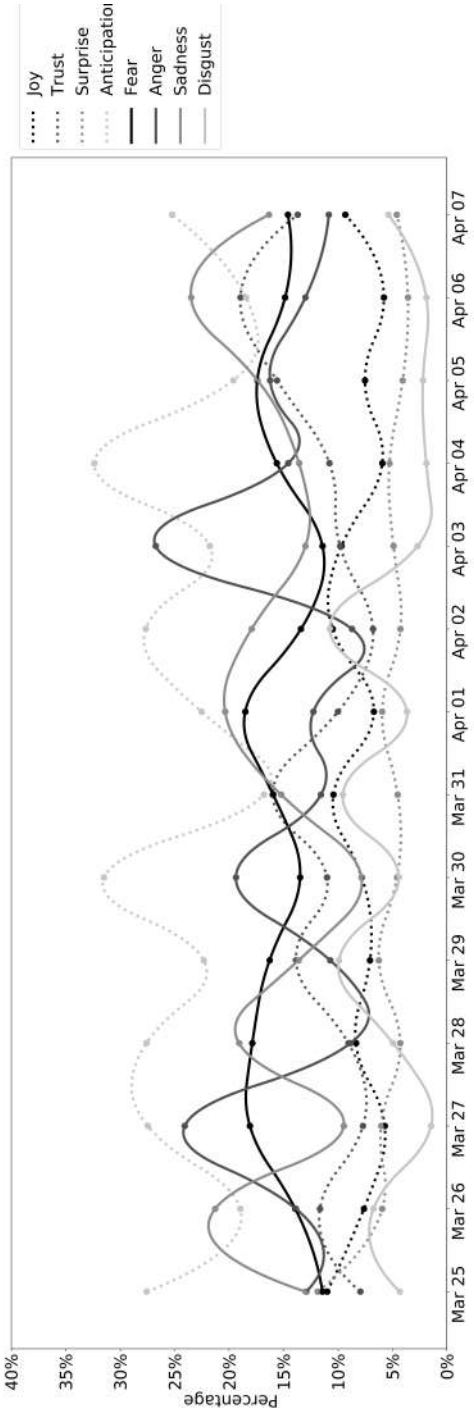

Fig. 6: Emotion trend on the word *mask* from March 25, 2020 to April 7, 2020.

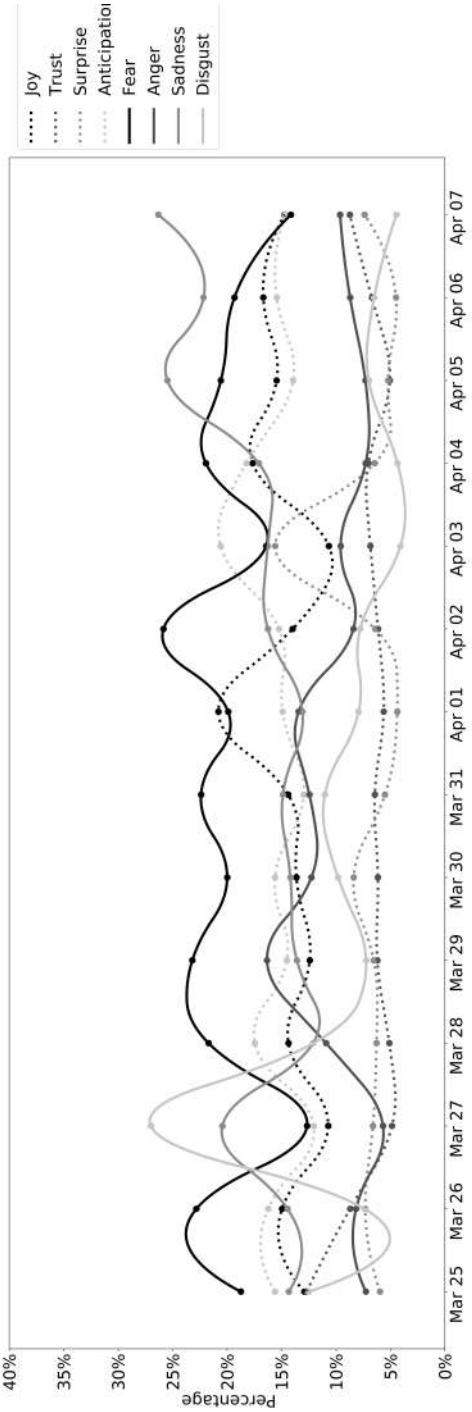

Fig. 7: EEmotion trend on the word *lockdown* from March 25, 2020 to April 7, 2020.

ID	Topic Label	Keywords
1	activities, life	people going everybody parties kickbacks majority really recover restaurant overwhelming spreads..
2	breaking affairs	murder covid-19 trump urgent doctors deaths lives lockdown first crisis..
3	individual health	covid19 lockdown care testing people time fever americans health work..
4	public health	cases deaths like county months died lockdown health know would state death next breaking..
5	politics	lockdown americans know left decided tired country house playbook destroy 69-page misleading crowd fight..

Table 4: Topics learned using LDA model: the `Topic Label` are manually summarized from the learned top `Keywords`.

and 9 in 10 Americans were staying home; India and South Africa joined the countries to impose lockdowns. Given that the United States, India and Brazil have large group of twitter users, we assume that this dramatic change may be triggered by those news.

5.2 Emotion trend with topics

To understand the mental health from a population level, we compare and study the emotion trends with topics. We first use the latent Dirichlet allocation (LDA) [3] model to learn fine-grained topics. LDA is a probabilistic model for discrete data (in our case, the tweets) and can learn topic clusters in an unsupervised way, where each topic is represented by a list of keywords. We sampled ten days' data randomly from our collected data (March 24 - May 30, 2020) and tried with different sets of hyper-parameters of the LDA model. The table shows the five fine-grained topics and their corresponding representative keywords. We then manually labeled the topics showed in `Topic Label` column by summarizing from the keywords in the cluster.

We looked at a larger time period of eight weeks. We selected the data in from the following dates: (Early) March 25, March 29, April 1 and April 5; and (Later) May 6, May 9, May 13 and May 16. We basically chose two weeks from the early stage and two weeks from the later stage, then form each week, 2 days' data were selected randomly. Then the topics were inferred and emotion classification was conducted using the BERT(ft) model.

Fig 8 shows the emotion trend in the topic `politics`. There is a slight drop in the positive emotions including *joy* and *trust*; more people are feeling *fear* and *anger*, but less are feeling *disgust* and *sadness* in the later stage. Fig 9 shows the results for `individual health` where we see that an improvement happened in the joy emotion but still the top three emotions are *fear*, *anger* and *sadness*. In general, form the population level, we notice that many of the

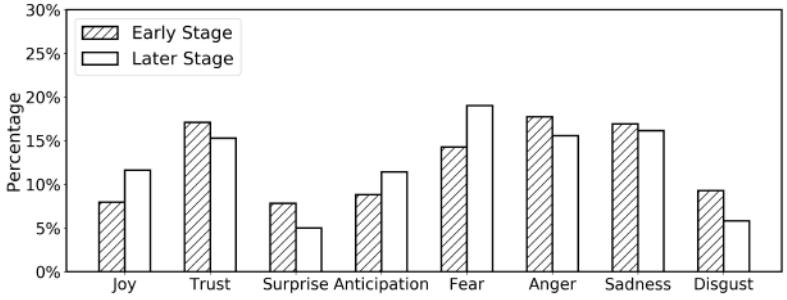

Fig. 8: Emotion trend on Politics.

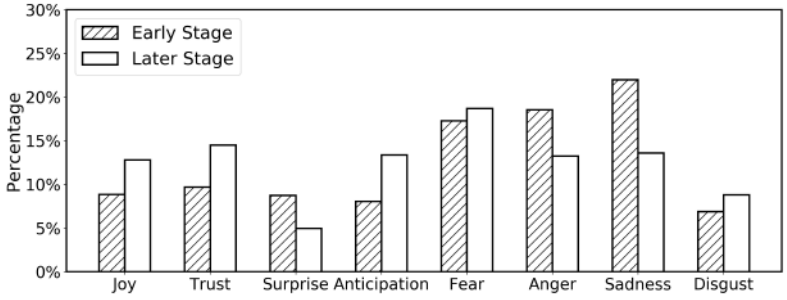

Fig. 9: Emotion trend on Individual Health.

tweets are classified as a negative feeling than positive ones. From our data and analysis, we conclude that the COVID-19 pandemic potentially have had impacts on mental health on many perspectives like health conditions and the society (i.e.,politics).

6 Conclusion and Future Work

In this work, we build the EmoCT dataset for classifying COVID-19-related tweets into different emotions to study the mental health problem. Based on this dataset, we conducted both single-label and multi-label classification tasks and achieved promising results. To understand the reasons why the public may feel sad or fear, we applied two methods to calculate correlations of the keywords and conducted some analysis to study the emotion trend. In the future work, we will study more in-depth analysis to better understand how COVID-19 affect on mental health. Besides, it will be helpful for us to have a correct estimates of the COVID-19 effects on people's long term mental health.

References

1. Abidin, T.F., Hasanuddin, M., Mutiawani, V.: N-grams based features for indonesian tweets classification problems. In: 2017 International Conference on Electrical Engineering and Informatics (ICELTICs). pp. 307–310. IEEE (2017)
2. Althoff, T., Clark, K., Leskovec, J.: Large-scale analysis of counseling conversations: An application of natural language processing to mental health. Transactions of the Association for Computational Linguistics **4**, 463–476 (2016)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
4. Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering **23**(5), 649–685 (2017)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dini, L., Bittar, A.: Emotion analysis on twitter: The hidden challenge. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3953–3958 (2016)
7. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision
8. Hasan, M., Rundensteiner, E., Agu, E.: Emotex: Detecting emotions in twitter messages (2014)
9. Larsen, M.E., Boonstra, T.W., Batterham, P.J., O'Dea, B., Paris, C., Christensen, H.: We feel: mapping emotion on twitter. IEEE journal of biomedical and health informatics **19**(4), 1246–1252 (2015)
10. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT) (2017)
11. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)
12. Pandey, A.C., Rajpoot, D.S., Saraswat, M.: Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management **53**(4), 764–779 (2017)
13. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages (2020)
14. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1524–1534. Association for Computational Linguistics (2011)
15. Zhang, M., Zhou, Z.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering **26**(8), 1819–1837 (2014)