

What are We Depressed about When We Talk about COVID19: Mental Health Analysis on Tweets Using Natural Language Processing

Irene Li¹, Yixin Li¹, Tianxiao Li¹,

Sergio Alvarez-Napagao², Dario Garcia-Gasulla² and Toyotaro Suzumura²

¹ Yale University, USA

² Barcelona Supercomputing Center (BSC), Spain

Yale University



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



Introduction



The outbreak of coronavirus disease 2019 (COVID-19) recently has affected human life to a great extent.

People talk about the pandemic on social media...

Introduction

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.

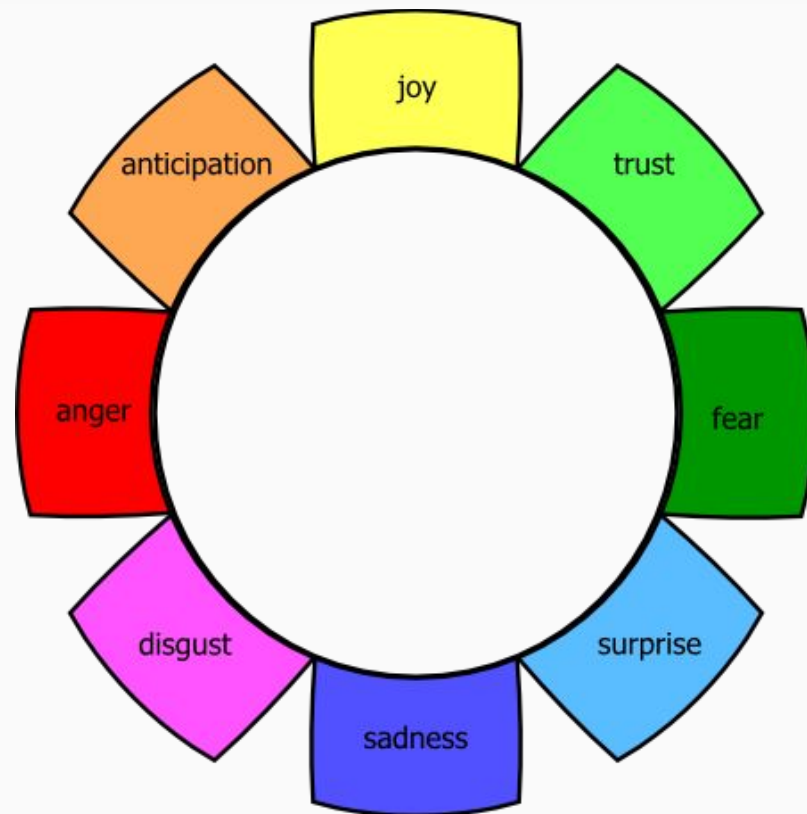


Contributions

Classification using deep models: anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

EmoCT (Emotion-Covid19-Tweet) dataset for the training purpose by manually labeling 1,000 English tweets.

Furthermore, we propose and compare two methods to find out the reasons that are causing sadness and fear.



Emotion Analysis

EmoLex (Mohammad and Turney, 2013; Hasan et al., 2014): NRC Word-Emotion Association Lexicon; PCAs, CNNs, LSTMs for sentiment classification. (Larsen et al., 2015, Alayba, Abdulaziz M., et al. 2018)

Pre-trained Language Modeling:

ELMo, GPT-2, GPT-3, **BERT**: contextualized word embeddings with attention. (Peters, Matthew E., et al. 2018, Radford et al., 2019, Devlin et al., 2019)

Twitter API, keyword list

Multiple languages.

Time span: March-May, 2020

A large capacity:

Each day ~ 3 million tweets

coronavirus, covid19, covid, COVID-19, covid 19, confinamiento, flu, virus, hantavirus, fever, cough, social distance, lockdown, pandemic, epidemic, conlabelious, infection, stayhome, corona, epidemie, epidemia, 新冠肺炎, 新型冠状病毒, 疫情, 新冠病毒, 感染, 新型コロナウイルス, コロナ

Data Collection: language distribution (March 24 to 26, 2020)

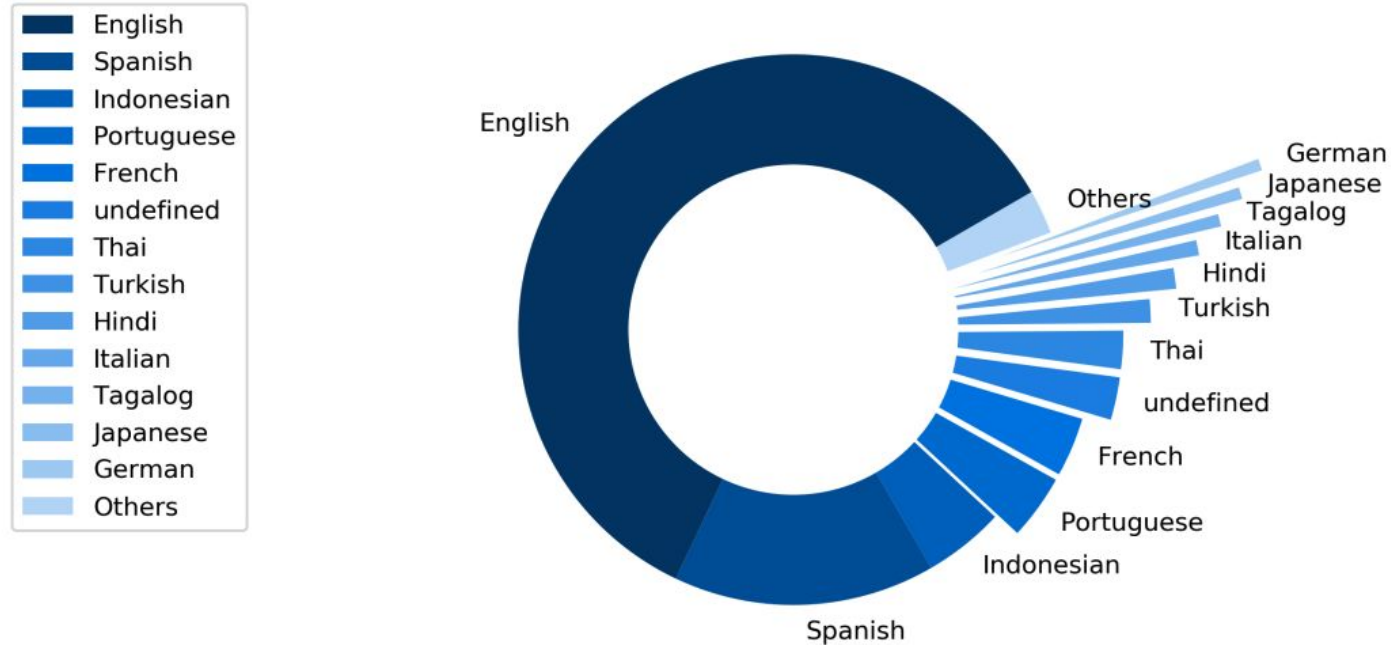


Figure 1: Language distribution on 8,148,202 tweets.

Data Collection: geolocation distribution (March 24 to 26, 2020)

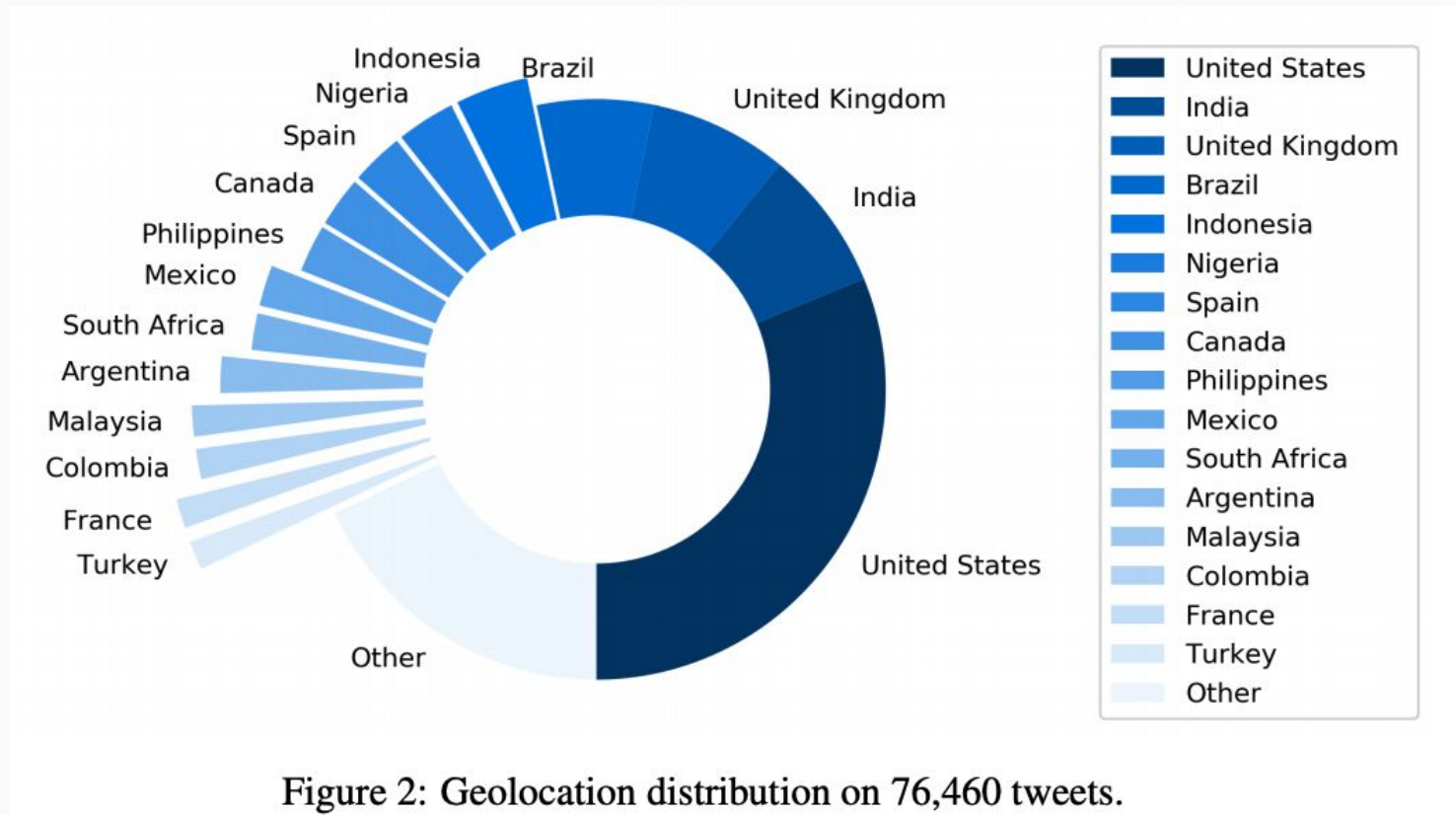


Figure 2: Geolocation distribution on 76,460 tweets.

EmoCT (Emotion-Covid19-Tweet) dataset

Selected covid-19 related English tweets: 1,000 (125 x 8)

Each tweet has up to 3 emotion tags.

For each emotion, we made sure that the primary label appeared in 125 tweets.

For each emotion:

Training 100; Testing 25

6905611270	4	1	7
3843569665	0	3	
3423583241	4		
6360377345	3		
2974462981	1	4	7

Emotion	tag code
Anger	0
Anticipation	1
Disgust	2
Fear	3
Joy	4
Sadness	5
Surprise	6
Trust	7

Sentiment Classification

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)

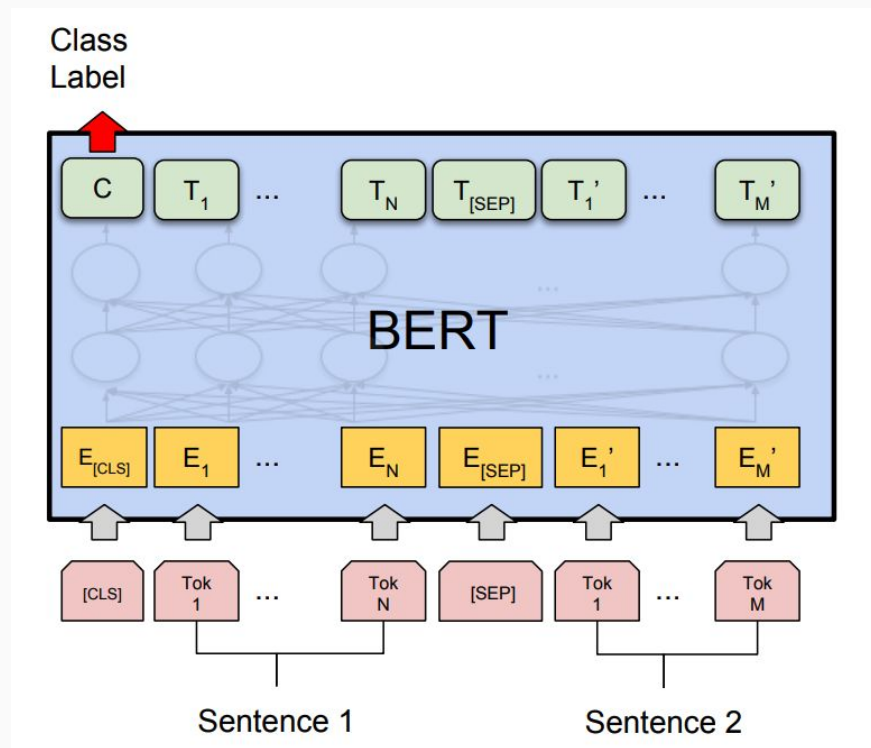
Multilingual Pre-trained BERT, then trained on our own dataset.

Single-label:

A single fully-connected layer, cross-entropy loss.

Multi-label:

A single-layer classifier, Sigmoid, which receives BERT output and predicts the possibility of containing each of the eight labels.



Single-label Classification

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.

Multi-label Classification

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.

Emotion Trend Analysis

Focus on two keywords: “Mask” and “Lockdown”

March 29th, 2020: overall 1 million tweets, “mask” 8,071, “lockdown” 31,146

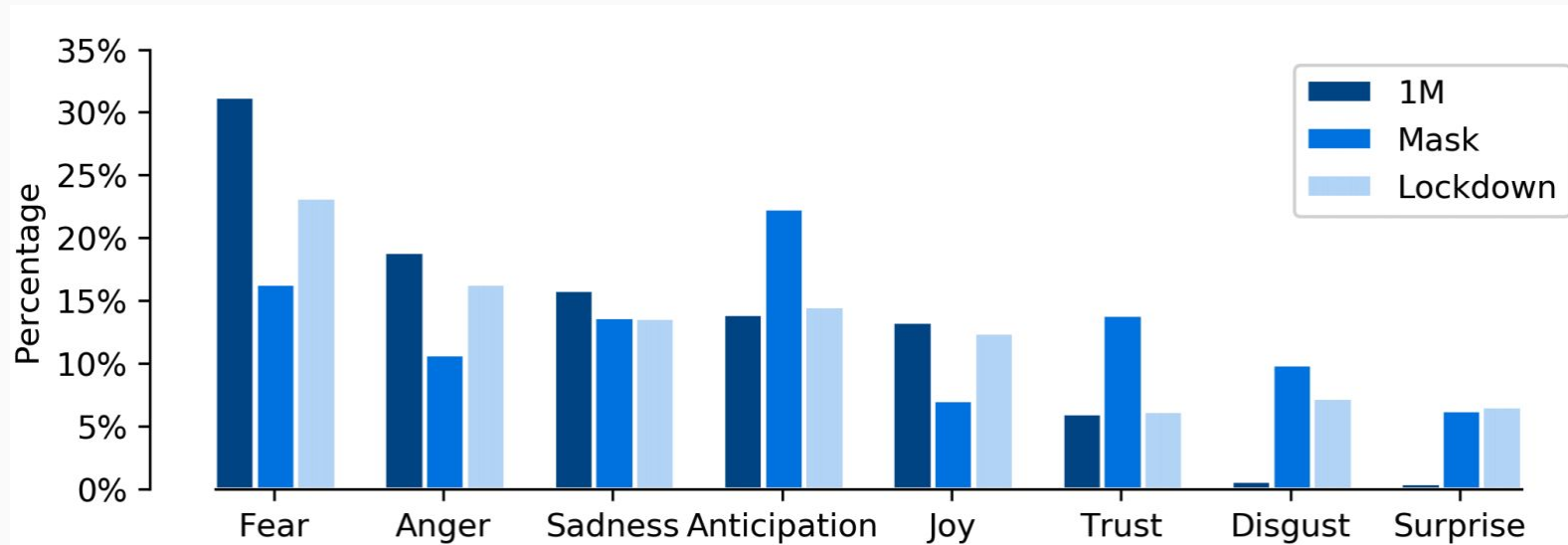


Figure 5: Comparison of emotion distribution.

Emotion Trend with keyword: "Mask"

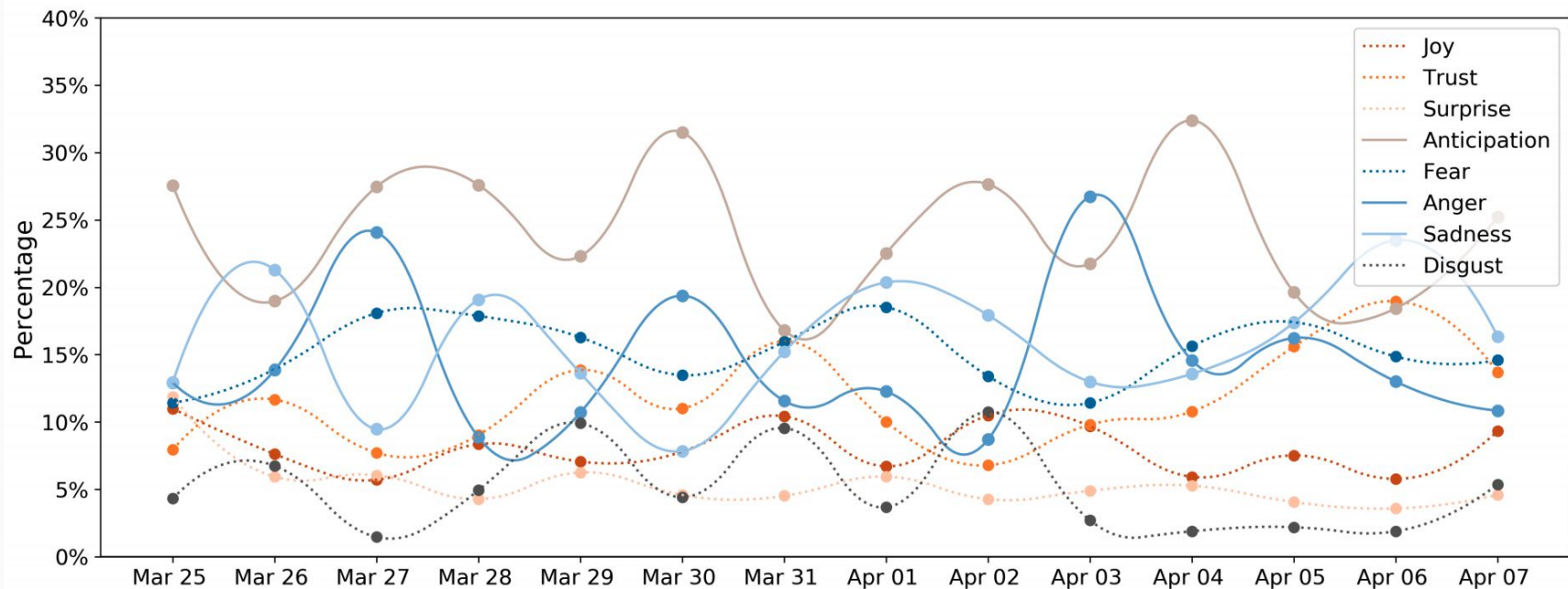


Figure 6: Emotion trend on the word *mask* from March 25 2020 to April 7, 2020.

Emotion Trend with Keyword: "Lockdown"

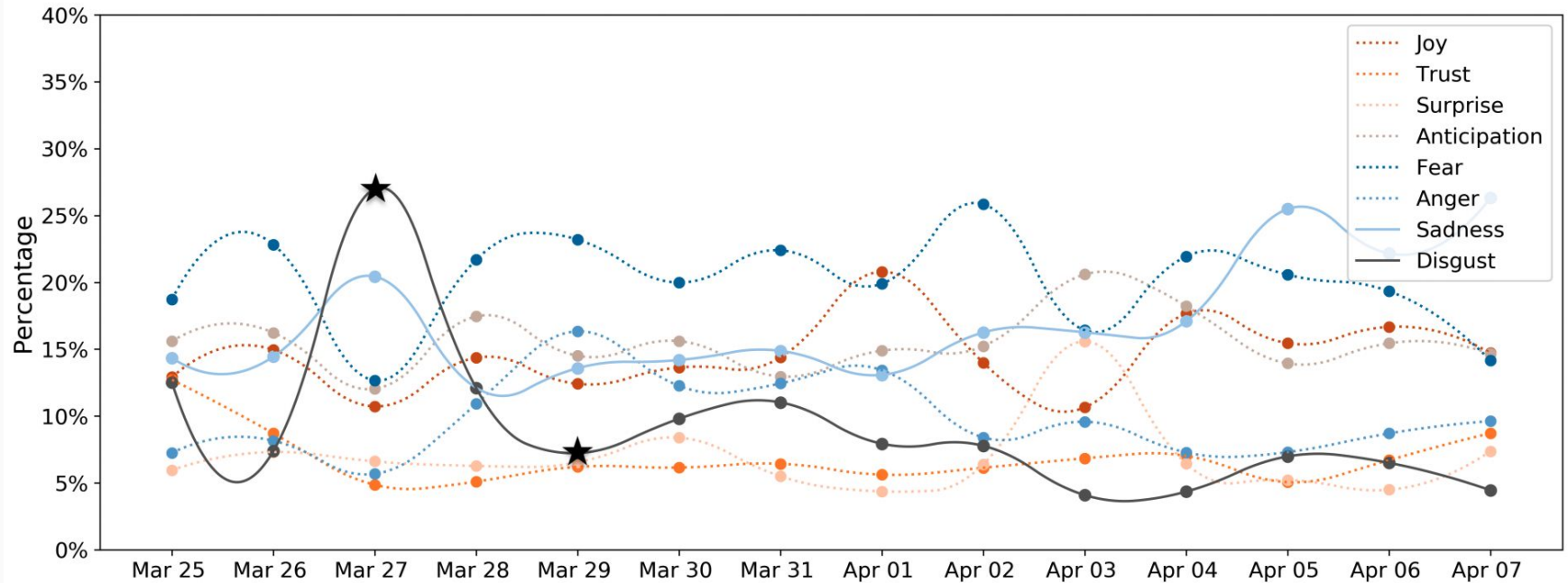


Figure 7: Emotion trend on the word *lockdown* from March 25 2020 to April 7, 2020.

Emotion Trend with Keyword: "Lockdown"

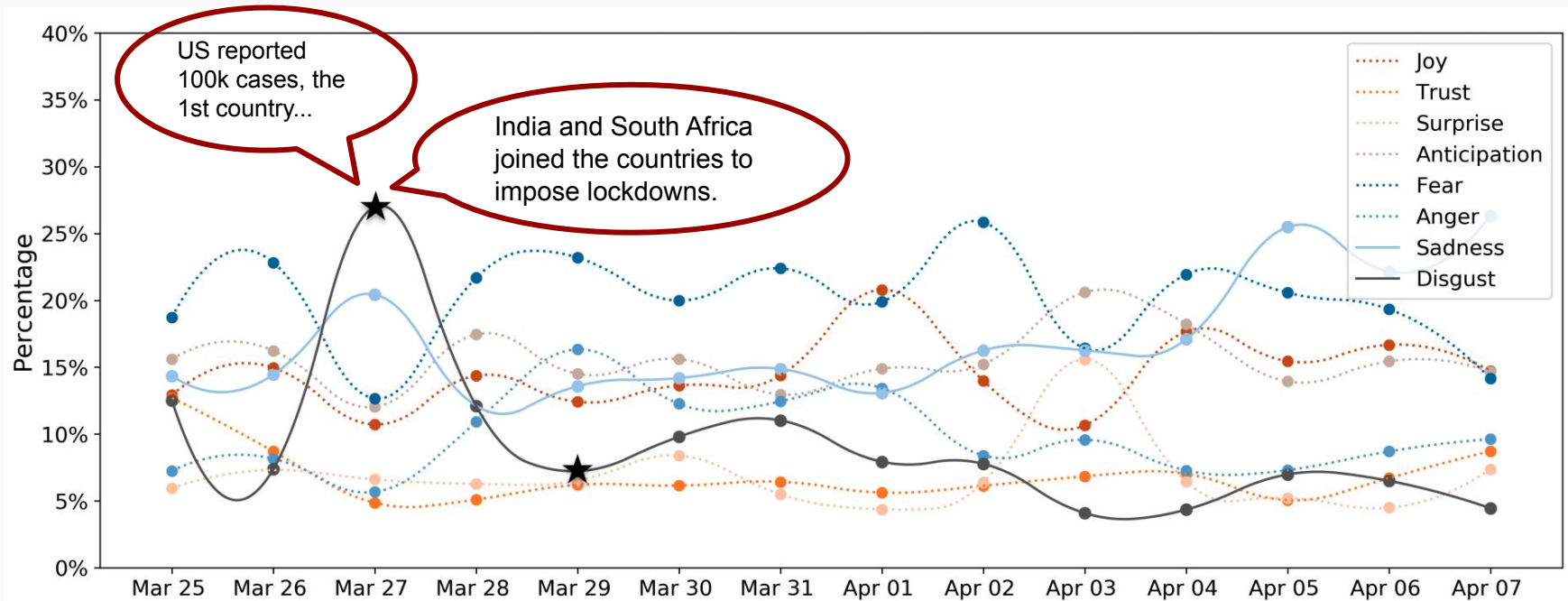


Figure 7: Emotion trend on the word *lockdown* from March 25 2020 to April 7, 2020.

Emotion Trend with Topics

Topic Modeling LDA (Blei et al. 2003):unsupervised machine learning method to cluster a set of documents into a number of clusters.

Sampled from 10 days **March 24 - May 30, 2020** to train an LDA topic model on 5 topics.
Show representing words for each topic:

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..

Emotion Trend with Topics

How the emotion trend would change?

(Early) late March
(Later) late May

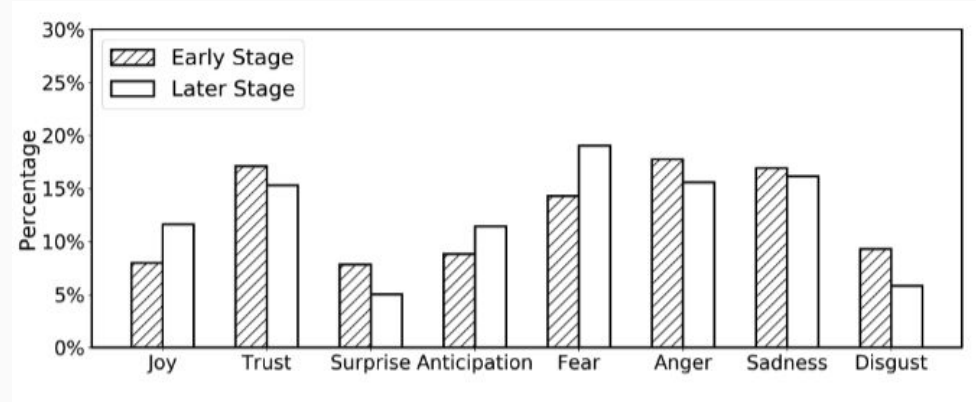


Fig. 8: Emotion comparison on **Politics**.

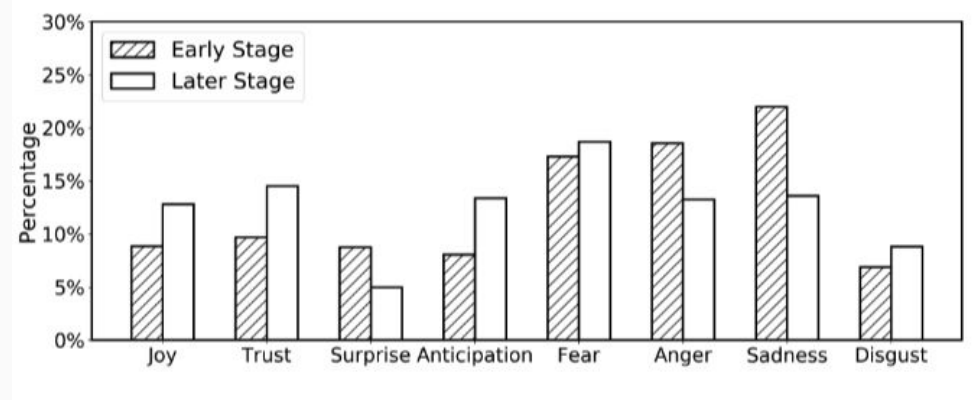


Fig. 9: Emotion comparison on **Individual Health**.

Emotion Trend with Topics

How the emotion trend would change?

(Early) late March
(Later) late May

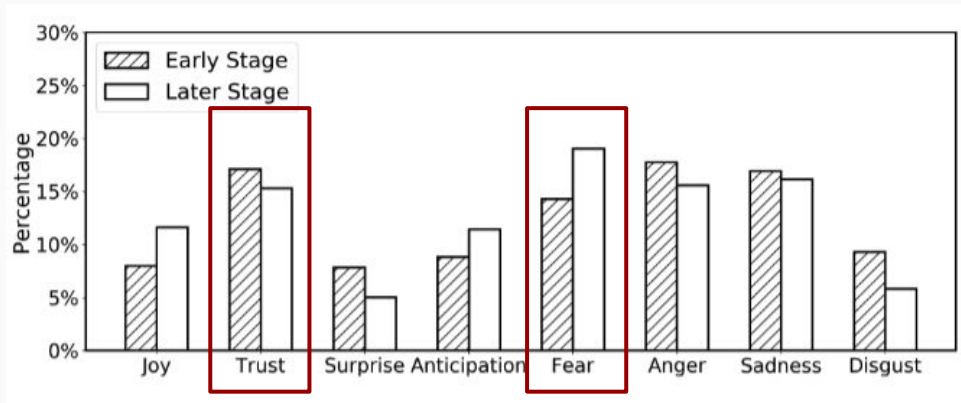


Fig. 8: Emotion comparison on **Politics**.

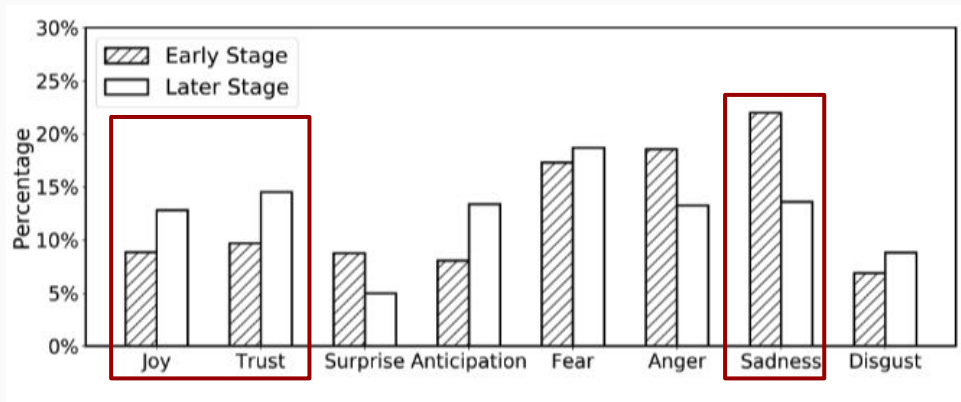


Fig. 9: Emotion comparison on **Individual Health**.

Conclusion and Future Work

We build the EmoCT dataset for classifying COVID-19-related tweets into different emotions to study the mental health problem.

Applied BERT on 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 keyword sand conducted some analysis to study the emotion trend.

Future Work: will study more in-depth analysis to better understand how COVID-19 affect on mental health, especially focusing on long-term trend analysis.

Thanks
Q&A