FineCOVIDSen: A Groundbreaking Fine-Grained Sentiment Analysis Dataset in COVID-19 Tweets

Anonymous ACL submission

Abstract

The COVID-19 pandemic had a profound global impact, necessitating a comprehensive understanding of public sentiment and reactions. Though there exist many public datasets about COVID-19, which advance in high volumes even reaching 100 billion, they suffer from the availability of labeled data or the coarse-grained sentiment labels. In this paper, we introduce FineCOVIDSen, a novel fine-grained sentiment analysis dataset tailored for COVID-19 tweets. It contains fine-grained ten categories varying in five different languages where each piece of data may contain more than one label. The dataset includes 10,000 annotated English tweets and 10,000 annotated Arabic tweets as well as 30, 000 translated Spanish, French, and Italian tweets from English tweets. Also, it comprises more than 105 million unlabeled tweets collected from March 1 to May 15, 2020. To support accurate fine-grained sentiment classification, we fine-tuned the pre-trained transformer-based language models on the labeled tweets. Beyond those, our study provides detailed analysis and unveils intriguing insights into the evolving emotional landscape over time in different languages, countries, and topics as well as a case study on the predicted results for unlabeled data. We also evaluate the availability of our dataset with ChatGPT. Our dataset and code are publicly available at anonymous GitHub [1]. Our hope is that this work will promote more fine-grand sentiment analysis on complex events for the NLP community.

1 Introduction

The global impact of COVID-19 has been profound, altering the lives of individuals worldwide. In order to curtail the transmission, measures such as quarantine, curfews, and social distancing have been widely implemented during this outbreak, leading to significant changes in work, education, and daily routines. Understanding people's reactions toward COVID-19 is crucial as it provides valuable insights into public perceptions and emotional responses toward the pandemic. By analyzing the sentiments expressed in social media, we can gauge the overall mood of the population, identify patterns of fear or anxiety, monitor public sentiment toward government actions and policies, and detect emerging concerns or issues (Lwin et al., 2020). This information is invaluable for policymakers, healthcare organizations, and researchers to make informed decisions, implement targeted interventions, and effectively address public concerns (Yue et al., 2019; Feng and Kirkley, 2021; Lazzini et al., 2022). Hence, it is essential to fulfill the sentiment analysis task for tracking global sentiments during the COVID-19 pandemic.

This task may initially appear straightforward given the extensive research on sentiment analysis in natural language processing (NLP) (Anees et al., 2020; Zhang et al., 2018; Kharde et al., 2016). However, it entails two significant challenges. **Firstly, it requires a substantial volume of tweets with sentiment annotations encompassing an extended time window following the outbreak.** To the best of our knowledge, there has not been any comprehensive dataset established for COVID-19 sentiment analysis with annotations on a large scale, as shown in Table 1. Take the recent dataset (Xue et al., 2020) for example, though it comprises 1.8 million tweets, it was not annotated and only analyzed through unsupervised methods based on topic modeling and lexicon features. **Secondly, tailored and fine-grained sentiment annotation labels are needed to better understand the impact of the health crisis**. Existing sentiment analysis tasks often utilize coarse-grained emotion labels such as "positive", "neutral", and "negative". However, the sentiments surrounding the pandemic are considerably more intricate compared to those encountered in mainstream sentiment analysis tasks. SemEval-2018 (Mohammad

[1] https://anonymous.4open.science/r/FineCovidSen-5F96

et al., 2018) is a tweet sentiment dataset comprising 11 categories. However, in the case of COVID-19, few tweets belong to *joy*, *love*, and *trust* categories, and numerous tweets from official sources were misclassified into inappropriate categories. Moreover, tweets containing jokes or denying conspiracy theories were not appropriately labeled. Based on our preliminary observation, the inclusion of adapted labels like *official report*, *joking*, *thankful*, and *denial* is indispensable for sentiment analysis in crisis-related tasks.

Herein, we are committed to developing **FineCOVIDSen**, a cutting-edge system powered by deep learning, designed specifically for tracking global sentiments during the COVID-19 pandemic. Our team diligently collected more than 105 million tweets related to COVID-19 encompassing five languages: English, Spanish, French, Arabic, and Italian. We annotated 10,000 tweets in English and 10,000 tweets in Arabic in 10 categories which are specifically designed for the pandemic, including *optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official*, and *joking*. We allowed one tweet to be annotated by more than one category, to support the multi-label analysis. We also translated the annotated English tweets into different languages (Spanish, Italian, and French) to augment our dataset for wide usage. We utilized a transformer-based framework to fine-tune pretrained language models on the labeled data and unveiled intriguing insights into the evolving emotional landscape over time in different countries and topics on the unlabeled data. Notably, we observed a gradual upsurge in optimistic and positive sentiments, which signifies a shared determination to surmount the obstacles presented by the pandemic and envisage a brighter future. This is consistent with the real case of COVID-19. We also demonstrate how our dataset proficiently mirrors public sentiment in relation to different parties and policies, proving to be a valuable tool for politicians during the stages of policy drafting and revision. Importantly, FineCOVIDSen offers a unique resource for various sentiment analysis tasks, which is valuable for the NLP community, especially on complex events that require fine-grained emotions.

The main contributions are summarized below:

a) We meticulously curated the largest fine-grained annotated dataset of COVID-19 tweets, comprising 10,000 English and 10, 000 Arabic tweets, annotated across 10 sentiment categories. This extensive dataset serves as a valuable resource for studying the social impact of COVID-19 and conducting fine-grained analysis tasks within the research community.

b) We provide a substantial collection of COVID-19 tweet IDs, meticulously collected since March 1, 2020, in five languages. This dataset spans 105 million tweets and will be continuously updated, allowing researchers to access a rich source of real-time COVID-19 discourse.

c) We report the usability of the labeled COVID-19 tweets by first evaluating the performance of deep learning classifiers and then testing on the 105 million unlabeled tweets to monitor how the global emotions vary in concerned topics and report other interesting findings as well as the availability evaluation with ChatGPT.

2 Related work

Sentiment analysis is contextual mining of text that identifies and extracts subjective information in the source material of the wider public opinion behind certain topics (Wang and Wan, 2018; Fei et al., 2022). To give a comprehensive summary of the existing works, we first review a group of selected works on non-COVID-19 tweets, and then a group of works on COVID-19 tweets in Table 1.

The general (non-COVID-19) tweet sentiment analysis often considers only a few general classes or ordinal sentiment scores (Srivastava and Bhatia, 2013; Priyadarshana et al., 2015; Balikas et al., 2017). For example, Sharma et al. classified tweets of movie reviews into *positive* or *negative* (Sharma et al., 2020). Deriu et al. trained a 2-layer CNN and a random forest classifier (RFC) for three sentiments (Deriu et al., 2016). When targeting fine-grained sentiments, the most popular benchmark dataset for tweet sentiment analysis is SemEval-2018, which is used for sentimental prediction (Baziotis et al., 2018; Jabreel and Moreno, 2018), and gender and race biases prediction (Kiritchenko and Mohammad, 2018). It has 7745 tweets in English, 2863 in Spanish, and 2863 in Arabic, labeled by 11 categories. Unfortunately, we discovered that the used labels on SemEval-2018 are inadequate for COVID-19 sentiment analysis. Specifically, we encountered a scarcity of tweets categorized as "joy", "love", and "trust", while a significant number of tweets from official sources were incorrectly assigned to inappropriate categories. Generally, the

Table 1: Summary of recent work on tweets sentimental analysis (None indicates 'not used', NA is 'not available')

Type	Related work	# Tweets		Sentiment category	Used model/algorithm
		Labeled	Unlabeled		
Non-COVID-19	(Deriu et al., 2016)	18K	28K	3 (positive, neutral, negative)	CNN+RFC
	(Baziotis et al., 2017)	61K	330M	3 (positive, neutral, negative)	LSTM+Attention
	(Mohammad et al., 2018)	15K	7,631	11 (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust)	Sentence embeddings + lexicons features
COVID-19	(Kabir et al., 2020)	None	700GB	3 (positive, neutral, negative)	Topic model (LDA)
	(Xue et al., 2020)	None	1.8M	8 (anger, anticipation, fear, surprise, sadness, joy, disgust, trust)	LDA + NRC Lexicon
	(Drias and Drias, 2020)	None	65K	10 (anger, anticipation, disgust, fear, joy, negative ,positive, sadness, surprise, trust)	Lexicon-based features
	(Kleinberg et al., 2020)	5K	None	8 (anger, anticipation, fear, surprise, sadness, joy, disgust, trust)	TF-IDF + POS features
	(Chen et al., 2020)	2M	None	2 (neutral, controversial)	LDA+sentimental dictionary
	(Barkur and Vibha, 2020)	None	24K	10 (anger, anticipation, disgust, fear, joy, negative ,positive, sadness, surprise, trust)	Lexicon-based features
	(Alhajji et al., 2020)	58K	20K	2 (positive, negative)	Naïve Bayes
	(Sri Manasa Venigalla et al., 2020)	None	86K	6 (anger, disgust, fear, happiness, sadness, surprise)	Emotion dictionary
	(Ziems et al., 2020)	2.4K	30K	3 (hate, counter-hate, neutral)	Logistic regression classifier
	(Naseem et al., 2021)	90K	None	3 (positive, neutral, negative)	BERT
	FineCOVIDSen (Ours)	20K	105M	10 (optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official report, joking)	BART

existing works on non-COVID-19 tweets face the problems of coarse-grained sentiments and inappropriate labels.

In the group of recent works on COVID-19 tweet sentiment analysis, Kabir et al. first built a real-time COVID-19 tweets analyzer to visualize topic modeling results in the USA with three sentiments (Kabir et al., 2020). As contemporaneous works, Xue et al. used LDA and NRC Lexicon on the English tweets to predict the (single) label of data where similar sentimental categories are used (Xue et al., 2020). Kleinberg et al. used linear regression models to predict the emotional values based on TF-IDF and part-of-speech (POS) features (Kleinberg et al., 2020). Alhajji et al. studied the Saudis' attitudes toward COVID-19 preventive measures with naïve Bayes models to predict three sentiments (Alhajji et al., 2020). Chakraborty et al. used TEXTBLOB and AFINN for capturing labels of data (Chakraborty et al., 2020). Chen et al. used sentiment features and topic modeling to reveal substantial differences between the use of controversial terms in COVID-19 tweets (Chen et al., 2020). Barkur et al. used a lexicon-based method to analyze the emotions on the nationwide lockdown of India due to COVID-19 (Barkur and Vibha, 2020). Ziems et al. used a logistic regression classifier with linguistic features, hashtags, and tweet embedding to identify anti-Asian hate and counter-hate text (Ziems et al., 2020). Although these methods advanced in large volumes, they suffered from coarse-grained sentiments or unavailable labeled data. Also, the labels captured based on emoji lexicons lack the evaluation process of data quality.

We conclude that supervised studies suffered from the scarcity of labeled data, and coarse-grained or inappropriate sentiment labels while the size and availability of the sentimental dictionary limited unsupervised methods.

3 Dataset Construction

3.1 Data Collection

We employed Twint [2], an open-source Twitter crawler to collect tweets, which offers flexibility by allowing users to specify parameters, including tweet language and time period. The unified query used across these languages included terms such as "COVID-19", "coronavirus", "COVID", "corona", etc. To collect tweets in different languages, we use the Twitter API by setting the field "lang". Note that retweets are included in our dataset since retweets often contain additional user-generated content in the form of comments or opinions, which can be valuable for sentiment analysis. To efficiently gather the data, we deployed 12 instances of Twint on a workstation equipped with 24 cores to download daily updates from March 1 to May 15, 2020. More data will be released for regular updates and maintenance. The collected tweets were then saved as JSON documents and consolidated into a shared medium for subsequent pre-processing.

3.2 Data Annotation

After collecting large volumes of unlabeled tweets, we performed sentiment annotation on a randomly selected subset of 10,000 English and 10,000 Arabic tweets. These two languages were selected

[2] https://github.com/twintproject/twint

based on their popularity, as English and Arabic are among the top five most widely used languages globally[3]. Then to determine the sentiment categories, we engaged several domain experts who carefully reviewed a subset of the collected tweets and referred to the SemEval-2018. After multiple rounds of discussions, we finalized a set of 10 labels that encompass the complex range of emotions observed during the pandemic. These labels include *optimistic* (representing hopeful, proud, and trusting emotions), *thankful* (expressing gratitude for efforts to combat the virus), *empathetic* (including prayers and compassionate sentiments), *pessimistic* (reflecting a sense of hopelessness), *anxious* (conveying fear and apprehension), *sad*, *annoyed* (expressing anger or frustration), *denial* (towards conspiracy theories), *official report*, and *joking* (irony or humor).

Our data was labeled by Lucidya [4] which is an AI-based company with rich experience in organizing data annotation projects. To ensure reliable annotations, we recruited over 50 experienced annotators, who were native speakers or fluent speakers and trained with example tweets with suggested categories to guide the annotation process. Each tweet was independently labeled by at least three annotators. We allowed multi-label annotation to capture the nuanced and complex emotions experienced during the pandemic. To assess the quality and agreement of the sentiment annotations, following (Mohammad et al., 2018), we calculated the average inter-rater agreement ι to evaluate the annotation reliability. The English annotations achieved an ι value of 0.904, while the Arabic annotations achieved an ι value of 0.931. These high values indicate a substantial level of agreement among the annotators. Additionally, we calculated the Kappa coefficient κ as 0.381 and 0.549 for English and Arabic annotations, respectively, indicating fair and moderate agreement[5].

Considering that the translation tools have been well developed, we translated the labeled English tweets into Spanish, French, and Italian with Google Translate to illustrate whether our classifiers can work well. There are three benefits of the translation: (1) It increased the diversity of the dataset benefiting from recognizing sentiment expressions in different linguistic and cultural contexts. (2) It was a scalable way to create a larger training dataset without the need for manual labeling. (3) It was a cost-effective alternative to leverage existing labeled data for multiple languages. To evaluate the quality of translation, we calculated the BLEU score by comparing A and A', where A' is translated back by A(En)->B(Es)->A'(En) taking the English and Spanish for example. The BLEU is 0.33 (note that the SOTA machine translation model has BLEU4 = 0.39 using a tied transformer ()), which verifies the good translation quality.

To ensure compliance with Twitter's Terms of Service and FAIR principles, the fetched data undergoes initial processing where any user-relevant information is removed. The tweet IDs for the unlabeled data and a limited number of tweet texts for the labeled data are saved and stored in the Git repository. The dataset is licensed under Apache-2.0 license, which allows for the sharing and adaptation of the dataset under certain conditions.

3.3 Data Description

3.3.1 Statistics of Unlabeled Tweets

We collected more than 105 million tweets related to COVID-19, spanning from March 1 to May 15, 2020, encompassing five languages: English, Spanish, French, Arabic, and Italian. The daily volume of collected tweets for each language is illustrated in Fig. 1. The statistical analysis reveals a consistent pattern across languages, characterized by a rapid increase in global conversation around COVID-19 and a gradual decline. English

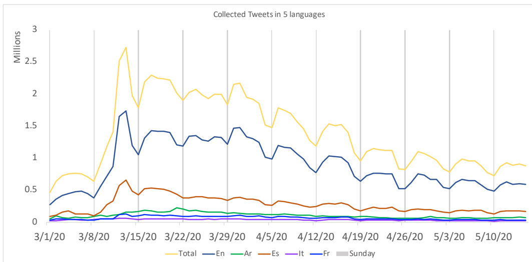


Figure 1: The absolute daily volume of COVID-19 Tweets collected in 5 languages, English (En), Spanish (Es), Arabic (Ar), French (Fr), and Italian (It). The vertical lines show Sundays, for guidance.

tweets dominate with the largest number and Spanish tweets take the second place followed by Arabic tweets, reaching the daily maximum on March 13 or March 21. In addition, people's attention cooled down as time went on. This trend was observed across different languages, suggesting that speakers

[3]https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets

[4]https://lucidya.com/

[5]https://en.wikipedia.org/wiki/Cohen%27s_kappa

Table 2: The label distributions of the annotated English, and Arabic datasets (%).

	Opti.	Than.	Empa.	Pess.	Anxi.	Sad	Anno.	Deni.	Offi.	Joki.
English	23.73	4.98	3.89	13.25	16.95	21.33	34.92	6.31	12.07	44.76
Arabic	11.27	3.33	6.49	4.65	7.53	10.80	17.17	2.10	34.52	14.18

of different languages responded to the pandemic in a similar manner. These features reflect the reliability and usability of our collected data.

3.3.2 Information of Annotated Tweets

The distribution of labels for each sentiment category in the annotated English and Arabic tweets is provided in Table 2. It is notable that the percentages do not sum up to 100% due to the multi-label annotation in our dataset. We find the difference in label distribution between English and Arabic tweets, this may lie in the different cultural backgrounds and religions. In English dataset, *joking* and *annoyed* emotions took large portions, which is consistent with the reality since COVID-19 causes deaths, high unemployment rates, and other problems. However, *optimistic* emotion represents the third largest category, indicating people also hold a sense of confidence and hope in combating the virus and envisioning a positive future. In Arabic dataset, the *official* label stands out significantly compared to the others, which was due to the numerous announcements and decisions made by Arabic governments in response to the outbreak. Table 3 (a) and (b) provide examples of English and Arabic tweets, demonstrating that some tweets exhibit multiple labels. Based on the category statistics, in English tweets, over 70% have multiple labels, while in Arabic tweets, about 20% do the same. Further analysis can be found in Appendix A.

Table 3: (a) English tweets examples

Category	Examples
	Single label
Opti.	Nothing last forever, Corona Virus will Vanish this month. “Happy New Month”
Than.	Gratitude to those who are involved to safeguard our lives from fatal Coronavirus. Thanks to them.
Anxi.	I don’t feel good and I don’t know if I’m just exhausted from working so much or if I have corona
Joki.	Calling Corona Virus “rona” like she the nastiest little girl in the 5th grade.
	Multiple labels
Pess., Joki.	if I get curved ima going somewhere packed to give myself coronavirus
Anxi., Pess.	Does everyone realize we’re going to reach a million cases of this coronavirus by the weekend?
Deni., Sad, Anno.	Why is it that no one ever reports on the number of people who recovered from Coronavirus?

(b) Arabic tweets examples

Category	Examples
	Single label
Opti.	والبطل من يطبق ما يريحه من عادات حتى بعد كورونا.. وكل شخص ينام على الجنب الي يريحه. لعلها خير!
Empa.	يا ريت نخصص لو خمس دقايق كل يوم لندعي ربنا يخلصنا من هالبلاء العظيم ، كورونا شلتنا
Anxi.	فكينا منك مالنا خلق نزورك في السجن تعرفين كورونا وخايفين
Anno.	اللهم شل اطرافک اللهم اجعل كورونا تعبث بجسدك يامفسد يابن سته وستين كلب سوف يتم ابلاغ الجهات المختصه وسوف تحاسب عجل غير اجل
Joki.	عيونک فايروس كورونا ، و قلبي صيني مايتحمل
	Multiple labels
Anxi., Sad	صرنا نخاف من الامراض والشهور ونسينا الخوف من الله كانت المساجد تذكرنا اليوم الاعلام كله يخوفنا من كورونا
Opti., Empa.	ربنا الشافي، ازمة كورونا آزالت غطاء السلوفان الإنساني الرقيق من علي هذا الشعب

4 Sentiment Classification Model

4.1 Data Preprocessing

As raw tweets are often short, unstructured, informal, and noisy, the first step of sentiment analysis is to preprocess the data. In detail, we first removed URLs from the tweet because they do not contribute to the tweet analysis. Then, we remove emojis and emoticons like ☺ though they can express emotions well since we focused on the analysis of textual data. Next, we filtered out noisy symbols and texts, that cannot convey meaningful semantic or lexicon information, and may even hinder the model from learning, such as the retweet symbol “RT” and some special symbols including line breaks, tabs, and redundant blank characters. Unlike previous methods which also removed hashtags in tweets, we kept these hashtags since they often encapsulate the main theme or topic of the tweet, making it easier to understand the subject matter. Apart from that, we also conducted word tokenization, steaming, and tagging.

4.2 Multi-label Sentiment Classifier

We built our multi-label sentiment classifier based on the Transformer due to its success on diverse NLP tasks. We fine-tuned the language models to train the customized classifier where two MLP layers were used. Particularly, we used BART (Lewis et al., 2019) for English, AraBERT (Antoun et al., 2020) for Arabic, and BERT (Devlin et al., 2018) for Spanish, French, and Italian. We also compare our method with other baselines including Fasttex, CNN, LSTM, LSTM-CNN, CNN-LSTM, BERT, BERTTweet, and XLNet on the FineCOVIDSen dataset. The same MLP layers were used.

We first train and evaluate the separate sentiment classifiers on the labeled English and Arabic tweets by 5-fold cross-validation. The well-trained model was then used for predicting the sentiments of mil-

Table 4: (a) Overall validation on FineCovidSen with a standard deviation

	Accuracy	F1-Macro	F1-Micro	LRAP	Hamm.Loss
En	0.498±0.008	0.535±0.012	0.580±0.008	0.548±0.007	0.156±0.004
Ar	0.591±0.010	0.488±0.016	0.614±0.008	0.635±0.009	0.083±0.002
Sp	0.428±0.004	0.434±0.010	0.511±0.003	0.493±0.002	0.177±0.001
Fr	0.430±0.010	0.432±0.010	0.509±0.010	0.496±0.009	0.176±0.004
It	0.437±0.006	0.442±0.010	0.517±0.005	0.503±0.005	0.172±0.002

(b) Accuracy of each category on FineCovidSen with a standard deviation

	En	Ar	Sp	Fr	It
Opti.	0.441±0.012	0.418±0.025	0.329±0.011	0.319±0.013	0.333±0.007
Than.	0.290±0.020	0.425±0.038	0.183±0.028	0.167±0.021	0.166±0.025
Empa.	0.438±0.018	0.459±0.042	0.243±0.032	0.278±0.024	0.292±0.056
Pess.	0.194±0.022	0.116±0.039	0.101±0.024	0.094±0.016	0.101±0.010
Anxi.	0.309±0.021	0.222±0.033	0.219±0.015	0.216±0.025	0.229±0.008
Sad	0.309±0.018	0.254±0.020	0.250±0.010	0.241±0.014	0.233±0.022
Anno.	0.514±0.016	0.389±0.032	0.429±0.010	0.428±0.023	0.430±0.014
Deni.	0.249±0.023	0.116±0.051	0.150±0.014	0.141±0.008	0.166±0.023
Offi.	0.619±0.019	0.872±0.017	0.566±0.017	0.569±0.025	0.576±0.022
Joki.	0.559±0.022	0.358±0.027	0.514±0.019	0.516±0.012	0.522±0.023

(c) Comparison of all models on FineCovidSen

Models	Accuracy	F1-Macro	F1-Micro	LRAP	Hamm.Loss
Fastext	0.371	0.269	0.453	0.469	0.162
CNN	0.389	0.387	0.482	0.470	0.178
LSTM	0.328	0.369	0.419	0.399	0.231
LSTM-CNN	0.312	0.380	0.413	0.368	0.264
CNN-LSTM	0.361	0.411	0.453	0.430	0.207
BERT	0.479	0.506	0.571	0.530	0.159
BERTTweet	0.498	0.535	0.585	0.542	0.159
XLNet	0.495	0.517	0.573	0.535	0.153
BART	0.498	0.535	0.580	0.548	0.156

Table 5: Performance Evaluation of Zero- and Few-shot Text Classification with ChatGPT on English Dataset

	Accuracy	F1-Macro	F1-Micro	LRAP	Hamm.Loss
Zero-shot	0.137	0.238	0.275	0.377	0.212
Few-shot	0.190	0.309	0.386	0.430	0.200

lions of COVID-19 tweets for our analysis.

4.3 Experimental Setting and Evaluation Metrics

We ran the experiments on a workstation with one GeForce GTX 1080 Ti. The batch size is 16, the learning rate is $4e-5$, and the models are trained in 20 epochs. The optimizer is Adam and the random seed is fixed as 42. We used multi-label accuracy, F1-macro, and F1-micro as well as ranking average precision score (LRAP) and Hamming loss to evaluate the performance.

5 Results and Analysis

5.1 Multi-label Classifier Validation

The performance evaluation of our sentiment classifiers for different languages on the FineCovidSen dataset is summarized in Table 4 (a). We find that the performance of the Arabic data is better than the English data. This is attributed to a higher rate of multiple labels in English tweets than in Arabic tweets. This proves that it is relatively challenging to classify English tweets. However, the accuracy of Spanish, French, and Italian tweets is worse than the original data. The reason is that the usage of different pre-trained language models: BART used for English tweets and AraBERT used for Arabic tweets perform better than BERT generally used for Spanish, French, and Italian on the same conditions (Yang et al., 2019; Antoun et al., 2020). It is worth noting that F1 values around 0.5 are influenced by the issue of class imbalance. The accuracy of each sentiment category in Table 4 (b) shows that *Official report*, *Joking*, *Optimistic*, and *Annoyed* can be predicted with an accuracy higher. *Pessimistic* and *Thankful* seem more difficult to predict than others. We illustrate the hot words of each category in Appendix C. We also compare some baselines in Table 4 (c). We see that BART performs almost best among all models followed by BERTTweet, XLNet, and BERT, which all belong to the group of Transformer. Fastext and CNN-LSTM have similar performance in that 1) Fastext has better power on OOV compared with Glove; 2) CNN better captures the local semantics compared with LSTM.

5.2 Availability Evaluation

To prove the availability of FineCovidSen, we feed our labeled data to GPT-3.5 for the multi-label text classification on the English data. We test them in the cases of zero-shot learning and few-shot learning on this task. As we can see in Table 5, the performance of the few-shot text classification is better than the zero-shot text classification on all metrics. This means that: 1) Our dataset is available in multi-label text classification; 2) It can be used for low-resource tasks with complex sentiments. More details are referred to in Appendix A.

5.3 Sentiment Variation

In this section, we present 1) **how sentiment varies in different languages**; 2) **how sentiment varies in different countries**; 3) **how sentiment varies in different topics**; 4) **how was the newly proposed emotion of Joking**; and 5) **how was public's attitude towards political parties.**

1) Sentiment Variation in Different Languages Over Days. We present the sentiment variation of the English tweets in Fig. 2. We see all positive emotions, including *optimistic, thankful* and *empathetic*, showed a similar trend of first rising up and then falling down. It implied people first felt positive due to the various decisions made for combating the virus in the middle of March. However, the emotions went down in late April when a large

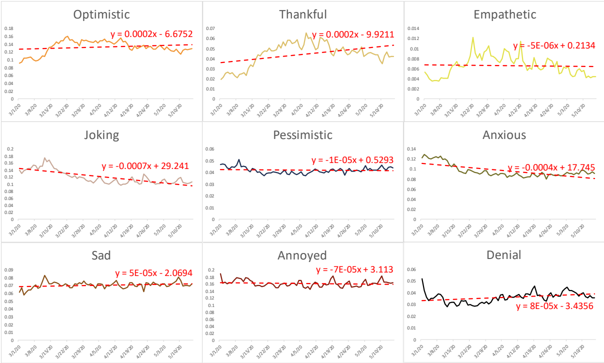


Figure 2: Sentiment variation of English tweets over time. The linear regression line of each emotion curve shows the trend of the emotion variation.

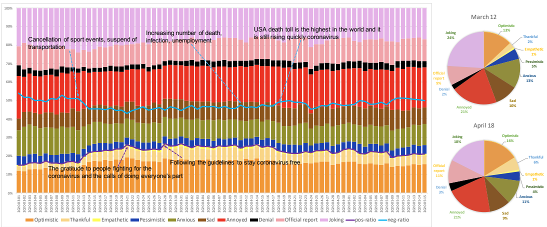


Figure 3: Sentiment variation in USA over time. Each bar shows the distribution of sentiments on one day (Better zoom in the spikes).

number of people got infected. Among negative emotions, *anxious* and *joking* fell down as time went on. The decrease of *anxious* may be caused by the increase in medical supplies. However, the high unemployment rate and death number may be the reason that *sad* and *annoyed* stayed high. The results of other languages are attached in Appendix B. *In summary, by examining how sentiment varies in different languages, we can gain insights into how people from diverse linguistic backgrounds express their opinions and emotions.*

2) Sentiments Variation of Different Countries Over Days. We selected the USA as an example to illustrate how the sentiments vary over days in Fig. 3. The blue and purple curves showed the positive (sum of *optimistic, thankful, empathetic* in yellow at different intensities) and the negative (sum of *pessimistic, anxious, sad, annoyed, denial* in blue at different intensities), respectively. We find that the portion of negative emotions was higher than that of positive emotions. On March 12, people felt *annoyed* and *anxious* (see the pie charts) since normal life was affected by the coronavirus e.g., cancellation of sports events and suspension of transportation. On March 21, however, the positive emotions had a slight increase when people were showing gratitude for the efforts of healthcare workers. The negative emotions went up once again due to the increasing rate of death, infection, and unemployment on April 11. The results of other countries are attached in Appendix B. *In summary, analyzing sentiment variations across countries helps identify regional sentiment trends, which were especially valuable for governments, healthcare organizations, and businesses to tailor their responses and communication strategies.*

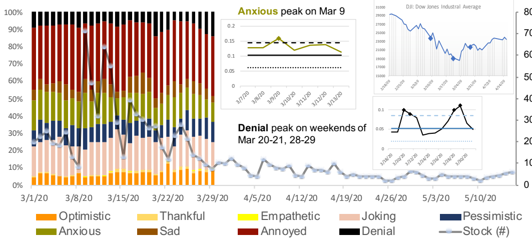
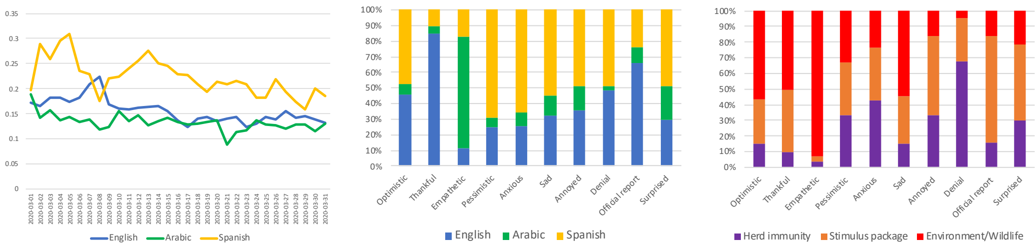


Figure 4: Sentiments variation on the stock market. We show the sentiment results when the topics were intensively discussed (around the peak of the volume curve in the background.

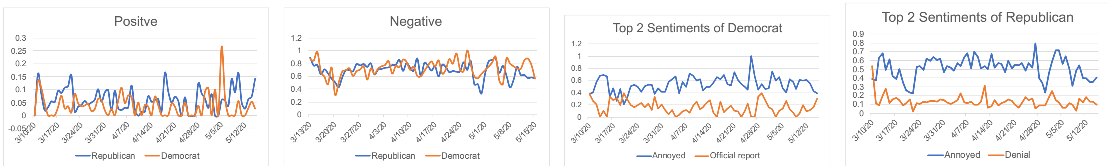
3) Sentiments Variation of Topics Over Days. We analyze the sentiment of the topic *stock market* in Fig. 4. It collapsed on March 9 when the peak of discussion was reached. *Anxious* reached a high value, which was greater than mean+2*std (out of the black dash line, and the black line is the mean, the dotted line is the mean-2*std). On March 12, the DJI (Dow Jones Index) had its worst day since 1987, plunging about 10% (the second time breakers) and the volumes arrived at the second largest. On the weekends of March 20-21 and March 28-29, the spikes of *denial* were higher than the blue dash line (mean+2*std), as a reflection of the continuous stock market collapse. The results of more topics are discussed in Appendix B, such as herd immunity, economic stimulus, and drug/medicine/vaccine. *In summary, investigating how sentiment differed across various COVID-19-related topics can provide insights into which aspects of the pandemic were polarizing or emotionally charged. This information can guide public health campaigns and communication strategies.*

4) Analyzing the Newly Proposed Emotion of *Joking*. We select three languages and three topics to analyze the interesting emotion *Joking*, which we first proposed in this work. Fig. 5 (a) shows that the portion of *joking* (including *ridicule*) in Spanish was much higher than that in English and



(a) (b) (c)

Figure 5: Analysis of the category *joking*. (a) The portion of *joking* overtime in 3 languages. (b) and (c) show the co-occurrence of joking and other labels in 3 languages and 3 events, respectively.



(a) (b) (c) (d)

Figure 6: Analysis of public's attitude towards the two political parties. (a) and (b) are the trend of positive and negative sentiment, respectively. (c) and (d) show the top two sentiments over time for political parties, respectively.

Arabic, which is possibly related to cultures and religions. Fig. 5 (b) indicates that *joking* is often assigned with *thankful* in English , with *empathetic* in Arabic and with *pessimistic, anxious* in Spanish. In Fig. 5 (c), we see in herd immunity, *joking* largely co-occurs with *denial* , while in the stimulus package, jokes were made with *official* reports . When discussing the environment, *joking* and *empathetic* co-occur significantly.

5) Analyzing the Public's Attitude towards Two Political Parties. In Fig. 6 (a) and (b), we displayed the trends in positive and negative sentiments for two political parties in the U.S. Overall, the Republican party garnered more positive emotional support, while both parties were on par with negative sentiment. By analyzing tweets, we find that the Democratic party was supportive of multiple rounds of economic stimulus, increased government spending, and investment, as well as expanded unemployment and health insurance. The Republican party favored tax cuts and subsidized large corporations and hospitals. In Fig. 6 (c) and (d), we selected the top two sentiments for political parties. For the Republican party, the highest level of annoyance sentiment was registered on April 27, 2020, largely attributed to the postponement or outright denial of coronavirus relief measures. Similarly, denial sentiment reached its pinnacle on March 10, 2020, due to conflicts between President Trump and Democrats regarding a stimulus package. The Democrat party saw a spike in annoyance sentiment on April 26, 2020, which can be traced back to the GOP's insertion of $174 billion in tax breaks favoring the wealthy. *In summary, monitoring sentiment towards political parties over time can help gauge public opinion and track how political responses to the pandemic influence public sentiment. This can be valuable for political analysts, policymakers, and political parties.*

6 Conclusion

This paper presents the FineCovidSen, a fine-grained sentiment analysis benchmark dataset for COVID-19 tweets. The contributions include a large annotated data of 20,000 labeled English and Arabic tweets with 10 fine-grained categories, as well as 105 million unlabeled COVID-19 tweets in 5 languages. We fine-tune the Transformer-based models as the multi-label classifiers and apply the well-trained models to predict the labels of unlabeled tweets. We provide detailed analysis and unveil intriguing insights into the evolving emotional landscape over time in different languages, countries, and topics as well as a case study on the predictions. We employ ChatGPT on FineCovidSen to prove its availability on the zero- and few-shot settings. The FineCovidSen dataset offers a unique resource for various sentiment analysis tasks requiring fine-grained emotional analysis.

7 Limitations and Ethics

Limitations. Our dataset covers a limited number of tweets released from March 1, 2020, to May 15, 2020, compared to the BillionCOV (Lamsal et al., 2023) which was used for efficient hydration with more than billions of COVID-19 tweets. The sentiment analysis we did was during the outbreak, and we leave the research on post-COVID sentiment analysis for future work. Although we collected tweets in the top five languages, the sentiments expressed in other languages or specific regions might not be adequately represented. Additionally, the tweets collected from Twitter's API may not represent the entire population accurately, introducing potential biases in the sentiments expressed.
Ethics. In conducting sentiment analysis on social media data, it is important to consider ethical implications such as privacy, consent, and data protection. As we introduced in Section 3.3, we remove user-relevant information to comply with data privacy regulations. Besides, tweets can reflect biases in society, including but not limited to gender, race, and socioeconomic status, which are not considered when collecting and applying data in our work. For instance, when analyzing the public sentiments towards political parties, we do not tend to infer the political leanings of users but analyze people's sentiments towards political parties about the actions of COVID-19, such as stimulus packages, government spending, investment, unemployment, and health insurance. Our dataset should be used for research purposes only.
Discussion. The FineCovidSen dataset will promote more fine-grand sentiment analysis on complex events for the NLP community. Analyzing a large number of unlabeled data provides great information for policymakers, healthcare organizations, and researchers, who can make informed decisions, implement targeted interventions, and effectively address public concerns during global health crises. In addition, due to the imbalanced properties of labels in our dataset, it will be a good source to solve the label imbalance problem of the multi-label classification task on our dataset FineCovidSen.

References

Mohammed Alhajji, Abdullah Al Khalifah, Mohammed Aljubran, and Mohammed Alkhalifah. 2020. Sentiment analysis of tweets in saudi arabia regarding governmental preventive measures to contain covid-19.

Ansari Fatima Anees, Arsalaan Shaikh, Arbaz Shaikh, and Sufiyan Shaikh. 2020. Survey paper on sentiment analysis: Techniques and challenges. Technical report.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Georgios Balikas, Simon Moura, and Massih-Reza Amini. 2017. Multitask learning for fine-grained twitter sentiment analysis. In *Proc. of SIGIR*.

Gopalkrishna Barkur and Giridhar B Kamath Vibha. 2020. Sentiment analysis of nationwide lockdown due to covid 19 outbreak: Evidence from india. *Asian journal of psychiatry*.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proc. of SemEval*.

Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. 2020. Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*.

Long Chen, Hanjia Lyu, Tongyu Yang, Yu Wang, and Jiebo Luo. 2020. In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for covid-19. *arXiv preprint arXiv:2004.10225*.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proc. of SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Habiba H Drias and Yassine Drias. 2020. Mining twitter data on covid-19 for sentiment analysis and frequent patterns discovery. *medRxiv*.

Hao Fei, Chenliang Li, Donghong Ji, and Fei Li. 2022. Mutual disentanglement learning for joint fine-grained sentiment classification and controllable text generation. In *Proc. of SIGIR*.

Shihui Feng and Alec Kirkley. 2021. Integrating online and offline data for crisis management: Online geolocalized emotion, policy response, and local mobility during the covid crisis. *Scientific Reports*.

Mohammed Jabreel and Antonio Moreno. 2018. Eitaka at semeval-2018 task 1: An ensemble of n-channels convnet and xgboost regressors for emotion analysis of tweets. *arXiv preprint arXiv:1802.09233*.

Md Kabir, Sanjay Madria, et al. 2020. Coronavis: A real-time covid-19 tweets analyzer. *arXiv preprint arXiv:2004.13932*.

Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*.

Rabindra Lamsal, Maria Rodriguez Read, and Shanika Karunasekera. 2023. Billioncov: An enriched billion-scale collection of covid-19 tweets for efficient hydration. *Data in Brief*, 48:109229.

Arianna Lazzini, Simone Lazzini, Federica Balluchi, and Marco Mazza. 2022. Emotions, moods and hyperreality: social media and the stock market during the first phase of covid-19 pandemic. *Accounting, Auditing & Accountability Journal*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

May Oo Lwin, Jiahui Lu, Anita Sheldenkar, Peter Johannes Schulz, Wonsun Shin, Raj Gupta, and Yinping Yang. 2020. Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. *JMIR public health and surveillance*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proc. of SemEval*.

Usman Naseem, Imran Razzak, Matloob Khushi, Peter W Eklund, and Jinman Kim. 2021. Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*.

YHPP Priyadarshana, KIH Gunathunga, KKA Nipuni N Perera, L Ranathunga, PM Karunaratne, and TM Thanthriwatta. 2015. Sentiment analysis: Measuring sentiment strength of call centre conversations. In *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*.

Prerna Sharma, Kanika Gupta, Manvi Bareja, and Vinay Kumar Jain. 2020. Effective approach for sentiment analysis on movie reviews. In *Soft Computing: Theories and Applications*.

Akhila Sri Manasa Venigalla, Dheeraj Vagavolu, and Sridhar Chimalakonda. 2020. Mood of india during covid-19–an interactive web portal based on emotion analysis of twitter data. *arXiv*.

Ritesh Srivastava and Mahinder Pal Singh Bhatia. 2013. Quantifying modified opinion strength: A fuzzy inference system for sentiment analysis. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *Proc. of SIGIR*.

Jia Xue, Junxiang Chen, Chen Chen, ChengDa Zheng, and Tingshao Zhu. 2020. Machine learning on big data from twitter to understand public reactions to covid-19. *arXiv preprint arXiv:2005.08817*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.

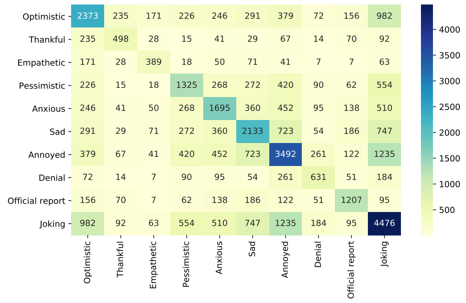
A Appendix: Dataset Information

1.1 Data Annotation

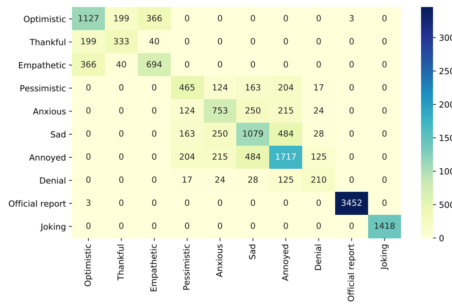
We introduce more details about data annotation. Our data was labeled by Lucidya which is an AI-based company with rich experience in organizing data annotation projects. The annotators were the native speakers or the fluent speakers. We allowed for multi-label annotation to capture the nuanced and complex emotions. Each annotator was trained with example tweets with suggested categories to

Table 6: Prompts for multi-label text classification

Zero-shot Prompt	Initialized: Multi-label Text Classification Model for Sentiment Analysis about COVID-19 Tweets. Instructions: This model classifies text inputs into different sentiments including "Optimistic", "Thankful", "Empathetic", "Pessimistic", "Anxious", "Sad", "Annoyed", "Denial", "Official report", and "Joking". Remember these three rules when making predictions: (1) Only use these ten sentiments for the predictions; (2) Each text may have more than one label; (3) Output all predictions of input texts.
Few-shot Prompt	Initialized: Multi-label Text Classification Model for Sentiment Analysis about COVID-19 Tweets. Instructions: This model classifies text inputs into different sentiments including "Optimistic", "Thankful", "Empathetic", "Pessimistic", "Anxious", "Sad", "Annoyed", "Denial", "Official report", and "Joking". Remember these three rules when making predictions: (1) Only use these ten sentiments for the predictions; (2) Each text may have more than one label; (3) Output all predictions of input texts. Examples:Input1: "Knowing I could've been taking in my new surroundings right now if it wasn't for Coronavirus ." "sentiment": "Sad, Joking" Input 2: "KAMALA HARRIS: Coronavirus treatment should be free BRIAHNA: ALL diseases matter!!" "sentiment": "Official report" ...



(a) English tweets



(b) Arabic tweets

Figure 7: Heatmaps of labels co-occurrence for English and Arabic tweets.

guide the annotation process. Each tweet was independently labeled by at least three annotators and paid 0.6 US dollars. The notebook of annotation guidelines is attached in the Supplementary Material.

To reduce the cheating cases during the annotation, we followed the below strategies: 1) The randomly selected small examples (50 pieces) were annotated by domain experts and our team members, and then provided to the annotation company. 2) Each annotator was trained in advance and must follow the annotation guidelines before he/she started to reach the full data. We used the small examples to train annotators and only the annotators who had a good performance (80% annotation accuracy) could participate in the annotation. 3) We regularly monitored annotators' performance and the quality of annotations. We allowed annotators to provide feedback and discuss with our domain experts about the labeled tweets with high uncertainty. Doing so allows us to select high-quality annotators for our multi-label annotation task.

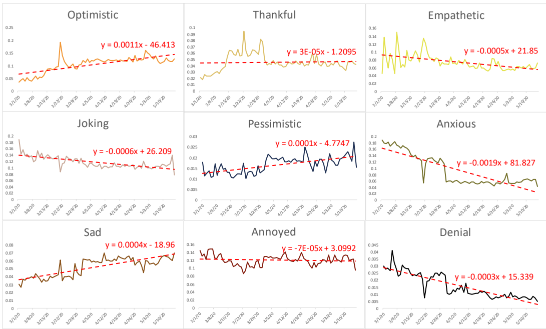
1.2 Label Co-occurrence of English and Arabic Data

To visualize the relationships between these labels in the English and Arabic data, we present the label co-occurrence heatmaps in Fig. 7. As shown in Fig. 7 (a), we see that the label co-occurrence is complex, which highlights the challenge of multi-label classification in the English dataset. In Fig. 7 (b), we see that the sentiment *Official* takes a large proportion compared to others, which results from that a lot of decisions were taken by the Saudi government.

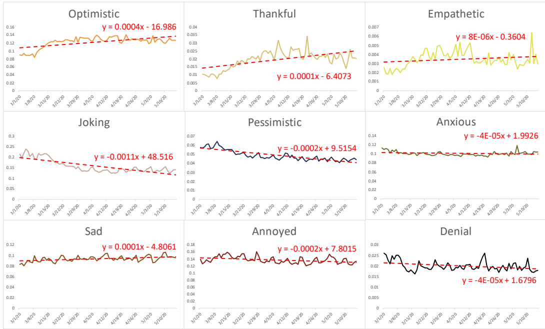
1.3 Label Distribution Variance

Based on the observation of labeled data and unlabeled data, one of the possible reasons is the different cultural backgrounds. On one hand, for the labeled data, the rate of the label "joking" is higher in English tweets than in Arabic while the rate of the label "Empathetic" in English is lower than in Arabic. On the other hand, for the unlabeled data, the predictions on them indicate the rate of the label "joking" shows a similar trend among English, Arabic, and Spanish where Spanish accounts for the first place, English is second place, and Arabic takes the last place. Therefore, this may be attributed to the intrinsic class imbalance.

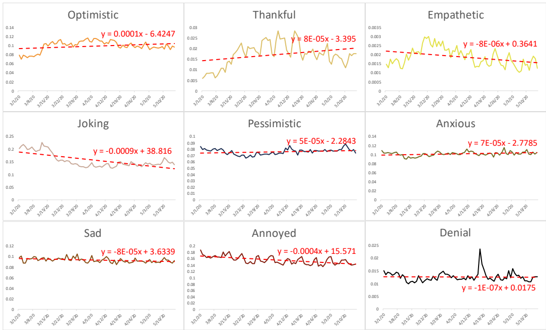
One more interesting phenomenon for the volume of daily tweets is that the number of tweets



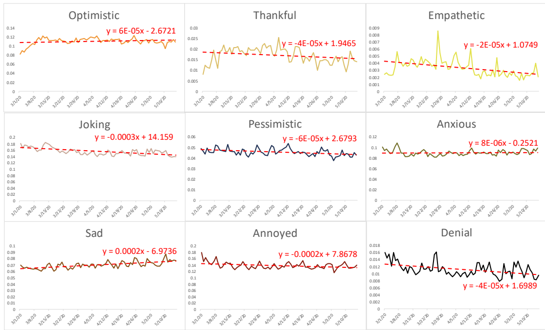
(a) Arabic



(b) Spanish



(c) French



(d) Italian

Figure 8: Sentiment variation of another four languages over time. Each subfigure corresponds to one type of language where nine emotions are reported. The linear regression line is fit to each emotion curve, showing the trend of the emotion variation.

shows a drop trend on Sunday as shown in Fig. 1. The possible reason is that Sundays are typically the weekend in many cultures, and people may be in activities that do not involve as much social media usage, such as enjoying time with family and participating in leisure activities.

1.4 Dataset Availability Evaluation With ChatGPT

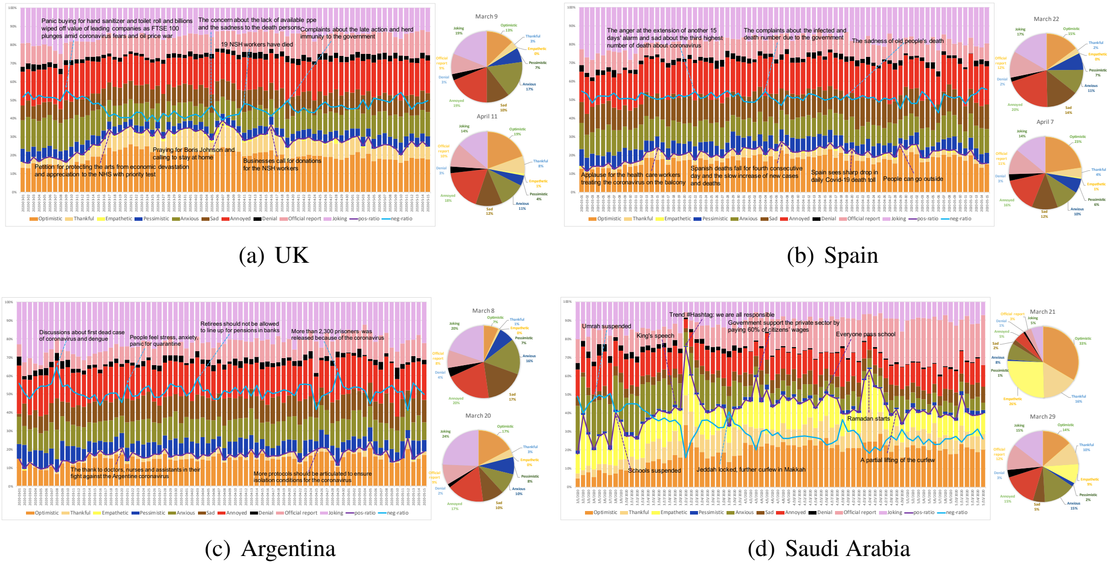
We run multi-label text classification using the labeled data on the zero-shot and few-shot settings on ChatGPT-3.5. For the zero-shot classification, we do not provide any labeled tweets to ChatGPT where only the prompt and label-removed data are fed. For the few-shot classification, we provide very limited labeled tweets to ChatGPT where only 38 out of 10, 000 tweets and the prompt are fed. Note that 38 tweets are randomly selected to ensure all of the labels can be seen by ChatGPT. The designed prompts are shown in Table 6.

B Appendix: More Interesting Findings of Sentiment Analysis

We present more analyzed results about sentiment variation including: 1) **how sentiment varies in different languages**; 2) **how sentiment varies in different countries**; and 3) **how sentiment varies in different topics**.

2.1 Sentiment Variation of Different Languages Over Days

The results of Arabic tweets shown in Fig. 8 (a) demonstrate significant variations in all categories of emotions. In particular, *optimistic* has been rising up, and *anxious, denial* and *joking* are falling down. The *sad* emotion keeps rising due to the increasing number of new cases in several Arabic-speaking populations, such as Saudi Arabia, Qatar, and the United Arab Emirates (UAE). The rise of *optimistic* and *thankful* and the fall of *pessimistic* and *annoyed* were also observed in Fig. 8 (b) of Spanish tweets. A similar trend of increase in *thankful* is observed in French tweets, as shown in Fig. 8 (c). However, the other emotions became stable, except the decline of *joking* and the sudden increase of *denial* to the conspiracy theory of the lab source of coronavirus. Italian tweets also showed a weak increase or decrease trends in most of the emotions, as shown in Fig. 8 (d), except those in *thankful* and *empathetic*.



(a) UK (b) Spain

(c) Argentina (d) Saudi Arabia

Figure 9: Sentiment variation in different countries over time. Each bar shows the distribution of sentiments on one day, where sentiments are shown in different colors. The blue curve and purple curve show the positive (sum of *optimistic, thankful, empathetic* in yellow at different intensities) and the negative (sum of *pessimistic, anxious, sad, annoyed, denial* in blue at different intensities), respectively. (Better zoom in to see the interpretation of spikes)

2.2 Sentiments Variation of Different Countries Over Days

Fig. 9 (a) showed in the UK, on March 9, the negative emotions caused by panic buying of hand sanitizer and toilet rolls and people's fear of coronavirus and oil price war leading to the plunging of the FTSE 100. After different coronavirus measures were imposed, the positive sentiment went up significantly. It would be better to zoom in on the figures to see other detailed interpretations.

In Spain (Fig. 9 (b)), people applauded the healthcare workers treating the coronavirus on the balcony on March 15, felt angry about the extension of another 15 days of alarm, and sad about the third highest number of deaths on March 22 (in the pie chart).

In Argentina (Fig. 9 (c)), the proportion of negative emotions was very close to 0.5 even much higher on some days. On March 8, the discussions about the first death case of coronavirus and dengue were focused on leading to the increase of *anxious*, *sad*, and *annoyed* (see pie chart at the right-hand). On March 21, the feelings of stress, anxiety, and panic went up because of the long quarantine, which resulted in the increase of *anxious* and *sad*. On April 29, more than 2,300 prisoners were released because of the coronavirus, which increased the feelings of *pessimistic*, *anxious*, and *annoyed*.

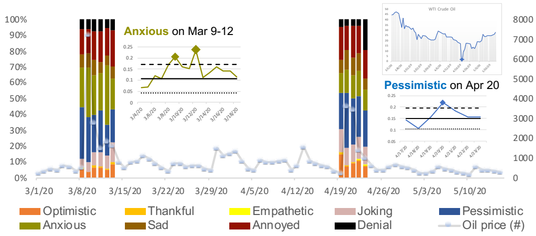
Fig. 9 (d) showed stronger positive sentiment in Saudi Arabia than in other countries or areas. Especially, starting from March 13, there was an increase in positive emotions when a lot of decisions were taken by the Saudi government. The peak was reached on March 21, responding to a tweet by the Saudi minister of health: "We are all responsible, staying home is our strongest weapon against the virus". Another positive peak was shown on April 23-24, when Ramadan started.

2.3 Sentiments Variation of Studied Topics Over Days

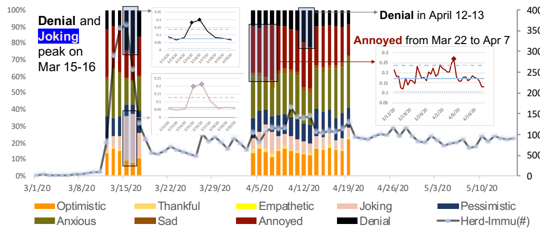
As shown in Fig. 10 (a), the topic of oil prices also showed the peak of discussion on March 9. The drop in crude oil price resulted in significant *anxious* on March 9-12. However, this was not the worst. On April 21, the crude oil price reached an 18-year low, which is shown on the marked point on the WTI crude oil curve. Among the triggered discussion, we see *pessimistic* was significant.

As shown in Fig. 10 (b), the topic of herd immunity quickly reached the top on March 14-15 when the UK government initially considered it on March 13. Among the intensive discussions from March 13 to 17, *denial* and *joking* were significantly observed on March 15-16. The discussion continued with significant *annoyed* from March 22 to April 7 and caused another rise of *denial* on April 12-13.

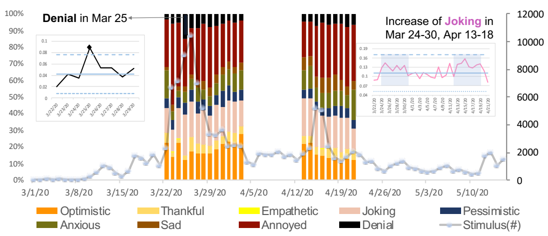
As illustrated in Fig. 10 (c), the topic of eco-



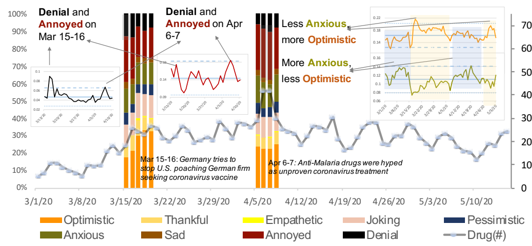
(a) Oil price



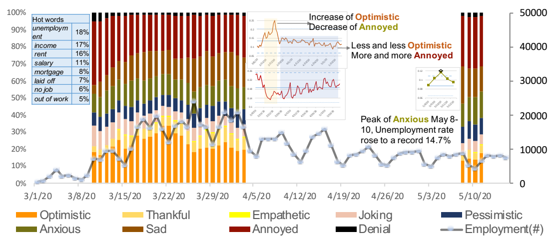
(b) Herd immunity



(c) Economic stimulus



(d) Drug/medicine and vaccine



(e) Employment/job

Figure 10: Sentiments variation on five topics. We show the sentiment results for these topics when they were intensively discussed (around the peak of the volume curve in the background).

nomic stimulus reached the top on March 26 when the US Senate passed a historic $2tn relief package. And another peak on April 15-16 when the checks were received. Surprisingly, during the discussion on March 23-26, positive was lower compared to other days, and *denial* was significant on March 25. We found many tweets under this topic, for example, “This is not enough”, “US economy is tanking”, and “The pandemic is getting worse”. By looking into the *joking*, we see increases on March 24-30 and April 13-18.

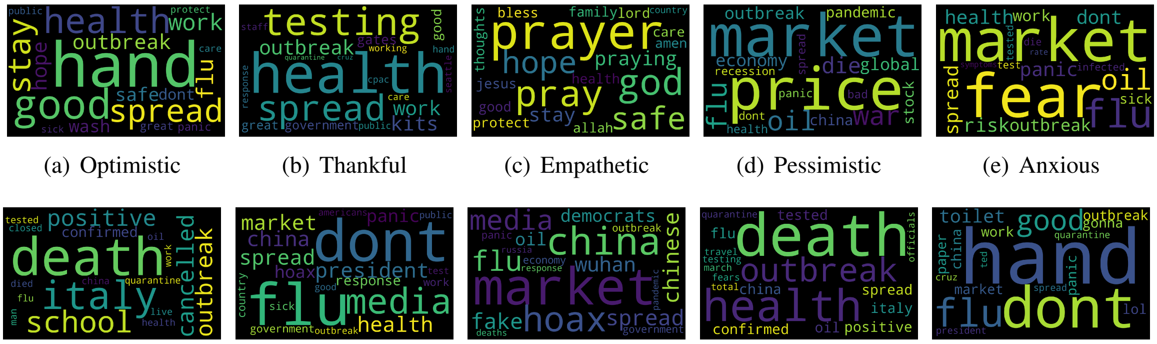
As we can see in Fig. 10 (d), the topic drug/medicine/vaccine collected the largest amount of discussion among these 5 topics (reaching 20-40K on the daily volume). This topic has been hot since the global outbreak around March 10. Two events caused significant *denial* and *annoyed*. The first event was on March 15-16, when Germany tried to stop the U.S. from poaching German firms seeking coronavirus vaccines. The second event was on April 6-7, when Anti-Malaria drugs were hyped as unproven coronavirus treatment. Overall from March to May, we see two sections of more *anxious* and less *optimistic*, and two other sections of less *anxious* and *optimistic*.

In Fig. 10 (e), the topic employment/job covered the hot words such as unemployment, income, rent, salary, mortgage, laid off, no job/work, etc. In March, we see an increase of *optimistic* and a decrease of *annoyed*, however, in April-May, we see less *optimistic* and an increase of *annoyed*. The peak of *anxious* was found on May 8-10, when the reported April unemployment rate rose to a record 14.7% in the US.

C Appendix: Hot Words Visualization

We present the hot words of the predicted English and Arabic tweets for each category where the date is randomly selected as March 9, 2020. The larger the word is, the more times it occurs in its category.

As we can see in Fig. 11, the class *optimistic* is represented by hand washing and health, which means people should wash their hands frequently to keep healthy. The class *thankful* is presented with Covid-19 testing, while the class *empathetic* is shown with “pray”, “hope", “god", and “safe".



(a) Optimistic (b) Thankful (c) Empathetic (d) Pessimistic (e) Anxious

(f) Sad (g) Annoyed (h) Denial (i) Official report (j) Joking

Figure 11: Hot words of each category for English tweets

(a) Optimistic (b) Thankful (c) Empathetic (d) Pessimistic (e) Anxious

(f) Sad (g) Annoyed (h) Denial (i) Official report (j) Joking

Figure 12: Hot words of each category for Arabic tweets

The class *pessimistic* is reflected in the economy market, oil market, and a large number of deaths. These hot words are also suitable for the class *anxious*. People felt *sad* about a lot of deaths and confirmed cases and the lockdown of schools. The class *annoyed* is displayed with “dont" and “flu" while the class *denial* is demonstrated with “market" and “China" since some people didn’t believe the Covid-19 report of China. Overall, these hot words in each category can represent the sentiments to some extent.

For Arabic tweets, we can see in Fig. 12 that the class *optimistic* is represented with الوقاية (protection), نمنع (prevent), علاج (treatment), and الخير (the good). The class *thankful* shows شكرا (Thanks), السعودية(Saudi), سلمان (Salman, the king), and الكـ ويت(Kuwait), which reflect how people are happy with governments actions against Covid-19. The *empathetic* words show the prayers to Allah for protecting the people and countries. The class *pessimistic* represents الناس (people), الشعب (commune), الحجر(quarantine), and أزمة(crisis). In *anxious* class, the words انتشار (spread), نخاف (fear), استغفر (asking forgiveness) are the popular words. The class *annoyed* represents المرض (disease), الصين (China), ايران (Iran), where the first case appeared in Saudi came from Iran. العالم (The world), حرب (war), مؤامرة (conspiracy) are the hot words in *denial* class which reflect how people think about this virus. The words in *joking* are الحجر (quarantine), البيت (house), الناس (people), and ابريل(April).