SenWave: A Fine-Grained Sentiment Analysis Dataset for COVID-19 Tweets

Anonymous ACL submission

Abstract

The global impact of the COVID-19 pandemic has prompted a need for a comprehensive understanding of public sentiment and reactions. Despite the existence of numerous public datasets on COVID-19, some reaching staggering volumes, even up to 100 billion, they face challenges related to the availability of labeled data and the presence of coarse-grained or inappropriate sentiment labels. In this paper, we present SenWave, a novel fine-grained sentiment analysis dataset tailored for COVID-19 tweets, covering ten fine-grained categories in five different languages. The dataset includes 10,000 annotated English tweets, 10,000 annotated Arabic tweets, and 30,000 translated Spanish, French, and Italian tweets from English tweets. Additionally, it encompasses more than 105 million unlabeled tweets from the first wave of COVID-19. To facilitate accurate fine-grained sentiment classification, we fine-tuned pre-trained transformer-based language models on the labeled tweets. Our study goes beyond this by offering detailed analysis and revealing intriguing insights into the evolving emotional landscape over time in different languages, countries, and topics. Furthermore, we evaluate the compatibility of our dataset with ChatGPT. Our dataset and code are publicly available on an anonymous GitHub [1]. We anticipate that this work will encourage more fine-grained sentiment analysis on complex events within the NLP community.

1 Introduction

The profound global impact of COVID-19 has led to significant changes in the lives of individuals worldwide. To mitigate transmission, various measures such as quarantine, curfews, and social distancing have been widely implemented, bringing about notable shifts in work, education, and daily routines. Analyzing sentiments expressed on social media provides a means to gauge the overall mood of the population. It enables the identification of patterns of fear or anxiety, monitoring of public sentiment toward government actions and policies, and detection of emerging concerns or issues (Lwin et al., 2020). This information holds immense value for policymakers, healthcare organizations, and researchers, facilitating informed decision-making, the implementation of targeted interventions, and effective addressing of public concerns (Yue et al., 2019; Feng and Kirkley, 2021; Lazzini et al., 2022). Therefore, understanding people's reactions to COVID-19 is crucial for gaining valuable insights into public perceptions and emotional responses to the pandemic.

Performing sentiment analysis during the COVID-19 pandemic presents significant challenges despite the abundance of research in natural language processing (NLP) on this topic (Anees et al., 2020; Zhang et al., 2018; Kharde et al., 2016). Two major challenges need to be addressed. **(1) Lack of Comprehensive Annotated Dataset.** Conducting sentiment analysis for COVID-19 requires a substantial volume of tweets with sentiment annotations, covering an extended time window following the outbreak. Despite various datasets, such as the recent one by (Xue et al., 2020) with 1.8 million tweets, no comprehensive dataset for COVID-19 sentiment analysis with large-scale annotations has been established (Table 1). Notably, existing datasets lack annotations and rely on unsupervised methods based on topic modeling and lexicon features. **(2) Lack of Tailored and Fine-Grained Sentiment Labels.** Unlike mainstream sentiment analysis tasks, sentiments surrounding the pandemic are intricate. Existing sentiment analysis typically employs coarse-grained emotion labels like "positive," "neutral," and "negative." However, these labels may not capture the complexity of sentiments during a health crisis. For instance, SemEval-2018, a tweet sentiment dataset with 11 categories, is not well-suited for COVID-19 sen-

[1] https://anonymous.4open.science/r/SenWave-73D0

timents. Categories like "joy", "love", and "trust" are underrepresented, and "official sources" tweets are misclassified. Additionally, tweets containing jokes or denying conspiracy theories lack appropriate labels. Therefore, incorporating adapted labels such as "official report", "joking", "thankful", and"denial" is crucial for effective sentiment analysis in crisis-related tasks.

Herein, we are committed to developing **SenWave**, a cutting-edge system powered by deep learning, designed specifically for tracking global sentiments during the COVID-19 pandemic. Our team diligently collected 105 million unlabeled tweets related to COVID-19 encompassing five languages: English, Spanish, French, Arabic, and Italian. We annotated 10,000 English tweets and 10,000 Arabic tweets in 10 categories including *optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official*, and *joking*. Also, we augment our dataset by translating the annotated English tweets into different languages (Spanish, Italian, and French) for wide usage. We utilized a transformer-based framework to fine-tune pre-trained language models on the labeled data and unveiled intriguing insights into the evolving emotional landscape overtime on the unlabeled data. The findings from the analysis revealed a steady increase in optimistic sentiments, aligning with the observed trends during the first wave of the COVID-19 pandemic. An interesting finding about the public sentiment of different parties and policies in the USA shows the value of our dataset on complex public events. Furthermore, we leverage ChatGPT to validate the efficacy of our dataset through zero-shot and few-shot multi-label sentiment analysis. Importantly, SenWave 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 conducted a thorough review of existing sentiment analysis datasets, identifying their limitations, which mainly revolved around either the absence of a comprehensive annotated dataset or the deficiency of tailored and fine-grained labels.

b) We diligently curated the most extensive fine-grained annotated dataset of COVID-19 tweets, featuring 10,000 English and 10,000 Arabic tweets annotated across 10 sentiment categories as well as 105 million unlabeled tweets. This comprehensive dataset stands as a valuable resource for exploring the social impact of COVID-19 and facilitating fine-grained analysis tasks within the research community.

c) We evaluate the effectiveness of the labeled tweets by first fine-tuning the Transformer-based models and second making predictions on the unlabeled data and finally analyzing the predicted results from different aspects. A ChatGPT-based evaluation of our dataset is done on the zero-shot and few-shot multi-label sentiment analysis.

2 Related work

2.1 Non-COVID-19 Tweets based Sentiment Analysis

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). Baziotis et al. used LSTM networks with attention mechanisms and pre-trained word embeddings on the tweets (Baziotis et al., 2017). When targeting fine-grained sentiments, the most popular benchmark dataset for tweet sentiment analysis is SemEval-2018 (Mohammad et al., 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, the used labels are inadequate for COVID-19 sentiment analysis since it encountered a scarcity of tweets categorized as "joy", and "love", while a significant number of tweets from official sources were incorrectly assigned such as "anticipation".*

2.2 COVID-19 Tweets based Sentiment Analysis

There are numerous public datasets on COVID-19 tweets (Kabir et al., 2020; Xue et al., 2020; Barkur and Vibha, 2020). For example, Kabir et al. built a real-time COVID-19 tweets analyzer to visualize topic modeling results in the USA with three sentiments (Kabir 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

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
	SenWave (Ours)	20K	105M	10 (optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official report, joking)	BART

predict three sentiments (Alhajji 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). 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 inappropriate labels or data quality evaluation.*

2.3 Aspect-based Sentiment Analysis (ABSA)

ABSA is the sentiemnt analysis task which not only focus on the sentiments but the aspects of input text while fine-grained SA works on a more granular level of labels. For example, Hoang et al. proposed to fine-tune BERT to a sentence pair classification model for ABSA (Hoang et al., 2019). Ma et al. proposed AMR-based Path Aggregation Relational Network for ABSA where the path aggregator and the relation-enhanced self-attention mechanism were used to efficiently exploit AMRs (Ma et al., 2023). Zhang proposed a SSEGCN architecture which integrated semantic information along with the syntactic structure for ABSA task by combining attention score matrices with syntactic mask matrices (Zhang et al., 2022).

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. Here, we focus on five kinds of languages including English, Spanish, French, Arabic, and Italian. The used terms across these languages in the query include "COVID-19", "COVID19", "coronavirus", "COVID", "corona", and corresponding Arabic ones. 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. The data were then saved as JSON documents for subsequent preprocessing. More data will be released for regular updates and maintenance.

3.2 Data Annotation

After collecting unlabeled tweets, we performed sentiment annotation on a randomly selected subset of 10,000 English and 10,000 Arabic tweets.

Sentiment Categories Determination. We enlisted the expertise of four domain experts with a rich background in public health and epidemiology. Experts first carefully reviewed a subset of the collected tweets then drew inspiration from SemEval-2018 and finally determined the ten sentiment categories 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),

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

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

anxious (conveying fear and apprehension), *sad*, *annoyed* (expressing anger or frustration), *denial* (towards conspiracy theories), *official report* (the release of factual information by governments or official organizations, such as confirmed cases, deaths, vaccine doses administered, and epidemic prevention policies), and *joking* (irony or humor).

Annotation Process. Our data was labeled by Lucidya [3] which is an AI-based company with rich experience in organizing data annotation projects. To ensure reliable annotations, we recruited 52 experienced annotators, who were native speakers or fluent speakers and trained with example tweets with suggested categories to guide the annotation process. Annotators referred to the annotation guideline notebook during annotation (See the anonymous GitHub [4]). Each tweet was independently labeled by three annotators. We allowed multi-label annotation to capture the nuanced and complex emotions experienced during the pandemic. The final labels of tweets were decided by the majority voting strategy when the annotated results were overlapped. Otherwise, the tweets were marked and reannotated to have consistent results. Finally, we got 10,000 annotated English tweets and 10,000 annotated Arabic tweets whose labels ranged in the mentioned ten categories.

Annotation Quality Evaluation. Followed by (Mohammad et al., 2018), we use Average Inter-rater Agreement (IRA) and Cohen's Kappa Coefficient to assess the quality and agreement of the sentiment annotations. IRA is measured as the average percentage of times each pair of annotators agree, while Cohen's kappa coefficient is a statistic to measure inter-rater reliability for qualitative items. The inter-annotator agreement scores are calculated by computing the scores three times for each pair of annotators, and then averaging these scores to obtain the final coefficient. Finally, the ι values for English and Arabic annotations reached 0.904 and 0.931 while the Kappa coefficients κ are 0.381 and 0.549, respectively. The high values of IRA indicate a substantial level of agreement among the annotators, and the good values of Cohen's kappa coefficient demonstrate fair and moderate agreement for our labeled data.

[3] https://lucidya.com/

[4] https://anonymous.4open.science/r/SenWave-73D0

3.3 Data Augmentation

Considering that the translation tools have been well developed, we translated the labeled English tweets into Spanish, French, and Italian with Google Translate to augment our dataset. There are three benefits of data augmentation: (1) It increases 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), verifying the good quality.

3.4 Data Overview

In this section, we provide an overview of the basic information regarding both the labeled and unlabeled tweets in our dataset.

3.4.1 Annotated Tweets

Data Distribution. The distribution of labels for each sentiment category in the annotated English and Arabic tweets is detailed in Table 2. In the English dataset, emotions such as *joking* and *annoyed* dominate, reflecting the harsh realities of COVID-19, including fatalities, high unemployment rates, and other challenges. Surprisingly, the *optimistic* emotion represents the third-largest category, suggesting that people also maintain confidence and hope in overcoming the virus and envisioning a positive future. In the Arabic dataset, the *official* label stands out significantly, attributed to numerous announcements and decisions made by Arabic governments in response to the outbreak. We observe differences in label distribution between English and Arabic tweets, potentially influenced by distinct cultural backgrounds and religions. In addition, the percentages of all labels do not sum

up to 100%, which is due to the multi-label annotation in our dataset. Also, the heatmaps of label co-occurrence for English and Arabic tweets are shown in Appendix A.

Data Examples. Table 3 (a) and (b) offer examples of annotated English and Arabic tweets. Analyzing the category statistics reveals that in English tweets, over 70% feature multiple labels, whereas in Arabic tweets, approximately 20% exhibit the same characteristic. Therefore, the sentiment analysis task on the English tweets is more challenging than Arabic, as shown in the experimental section.

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 اطرافك اللهم اجعل كورونا تعبث بجسدك يامفسد يابن سته وستين كلب سوف يتم ابلاغ الجهات المختصه وسوف تحاسب عجل غير اجل
Joki.	عيونك 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

3.4.2 Unlabeled Tweets

Data Volume. We collected 105 million of unlabeled tweets related to COVID-19, spanning from March 1 to May 15, 2020, covering the first wave of the pandemic. The tweets were gathered in five languages: English, Spanish, French, Arabic, and Italian. The daily volume of collected tweets for each language is depicted in Fig. 1. English tweets dominate with the largest number, followed by Spanish tweets, and then Arabic tweets, reaching their daily maximum on March 13 or March 21. These peaks coincide with significant events such as the US President declaring a national emergency, the Spanish Prime Minister declaring a state of emergency, and the Saudi Arabia suspending a public travel.

Data Glance. The trend of people's attention shows an initial increase to a peak point, followed by a gradual decline over time. This pattern is consistent across different languages, indicating a similar response to the pandemic among speakers of various languages. These characteristics underscore the reliability of our collected data. Interestingly, the number of tweets shows a drop trend on Sunday. 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.

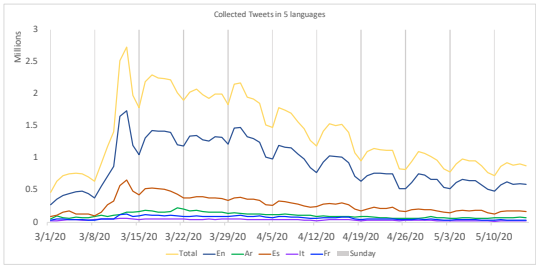


Figure 1: The absolute daily volume of COVID-19 Tweets collected in 5 languages, English (En), Spanish (Es), Arabic (Ar), French (Fr), and Italian (It). The vertical lines show Sundays, for guidance.

4 Sentiment Classification Models

4.1 Data Preprocessing

To prepare the raw tweets for sentiment analysis, we initiated the process with several preprocessing steps. Initially, we eliminated URLs as they do not contribute significantly to the sentiment analysis. Emojis and emoticons, such as ◡̈, were also removed, despite their expressive nature, as our focus was on analyzing textual data. Subsequently, we filtered out noisy symbols and texts that lack meaningful semantic, including the retweet symbol "RT" and special characters like line breaks, tabs, and redundant blank spaces. The user-relevant information is also removed to protect user privacy, such as usernames starting with the character @. Notably, unlike some prior methods, we retained hashtags in tweets, as they often encapsulate the primary theme or topic of the tweet, facilitating a better understanding of the subject matter. Additionally, we performed word tokenization, stemming, and tagging using the NLTK[5] for English, Spanish, French, and Italian, and Pyarabic[6] for Arabic.

[5]https://www.nltk.org/

[6]https://pypi.org/project/PyArabic/

Table 4: (a) Overall validation on the SenWave dataset

	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 the SenWave dataset

	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

4.2 Multi-label Sentiment Classifiers

Our multi-label sentiment classifier, rooted in the success of the Transformer architecture across various NLP tasks, was crafted by fine-tuning language models with a customized classifier featuring two MLP layers. Specifically, we leveraged 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. To evaluate the effectiveness of our approach, we compared it with several baselines, including Fasttex, CNN, LSTM, LSTM-CNN, CNN-LSTM, BERT, BERT-Tweet, and XLNet, all using the same classifier layers as ours. Non-Transformer-based methods employed 300-dimensional Glove embeddings for word representations. We use binary cross-entropy loss by averaging all labels as the loss function.

4.3 Experimental Settings and Evaluation Metrics

The experiments were conducted on a workstation equipped with one GeForce GTX 1080 Ti. The training setup included a batch size of 16, a learning rate of $4e-5$, and the models were trained over 20 epochs. The Adam optimizer was employed, and a fixed random seed of 42 was used for consistency. To assess performance, metrics such as multi-label accuracy, F1-macro, F1-micro, ranking average precision score (LRAP), and Hamming loss were employed. The evaluation was carried out using 5-fold cross-validation to ensure a robust assessment of model performance.

5 Results and Analysis

5.1 Multi-label Classifier Results

The evaluation results of the sentiment classifiers were summarized in Table 4 (a). We found the performance of the Arabic data outperformed that of the English data, which was attributed to a higher rate of multiple labels in English tweets compared to Arabic tweets. This suggested that classifying English tweets was relatively challenging. The accuracy of Spanish, French, and Italian tweets was worse than the original data. This was explained by the use of different pre-trained language models: BART and AraBERT perform better than the generally used BERT for Spanish, French, and Italian under the same conditions (Yang et al., 2019; Antoun et al., 2020). F1 values around 0.5 were influenced by the issue of class imbalance. The accuracy of each sentiment category in Table 4 (b) revealed that *official report*, *joking*, *optimistic*, and *annoyed* can be predicted with higher accuracy. On the other hand, *pessimistic* and *thankful* seemed more challenging to predict than others. In the comparison with baselines in Table 4 (c), BART performed almost the best among all models, followed by BERTTweet, XLNet, and BERT, which all belong to the Transformer group. Fastext and CNN-LSTM exhibited similar performance, where Fastext showed better out-of-vocabulary (OOV) capabilities compared to Glove, and CNN captures local semantics better than LSTM. The hotwords of each category of English and Arabic tweets are illustrated in Appendix C.

Table 5: Comparison of all models on the SenWave dataset

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

5.2 Dataset Reliability Evaluation

To validate the usability of SenWave, we employed GPT-3.5 for multi-label text classification on English data. We conducted tests in both zero-shot and few-shot learning scenarios. As shown in Table 6, the performance of few-shot text classification outperforms that of zero-shot classification across all metrics. This indicates two key findings: 1) Our dataset is effective for multi-label text classification; 2) It can be employed for low-resource tasks involving complex sentiments. The prompt of the multi-label text classification is in Appendix A.

5.3 Sentiment Variation of Unlabeled Tweets

In this section, we explore the variation of sentiments in different contexts, including **different**

Table 6: Zero-shot 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

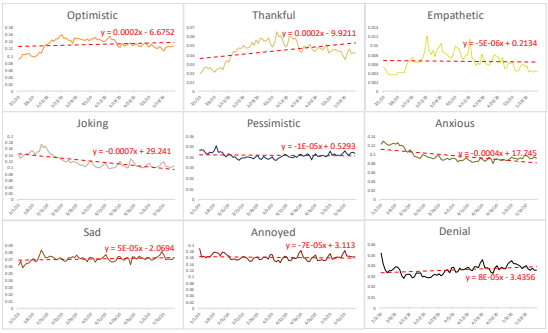


Figure 2: Sentiment variation of English tweets over time. The linear regression line of each emotion curve shows the trend of the emotion variation.

languages, **different countries**, **different topics**; and conduct the analysis on **the emotion of Joking** and **public attitudes towards political parties**.

1) Sentiment Variation in Different Languages Over Days. We illustrated the sentiment variation of English tweets in Fig. 2. All positive emotions exhibited a similar trend of initially rising and then declining. This suggests that people initially felt positive due to various decisions made to combat the virus in mid-March. However, these emotions declined in late April when a large number of people were infected. Among negative emotions, *anxious* and *joking* decreased over time. The decrease in *anxious* may be attributed to an increase in medical supplies, while the persistently high levels of *sad* and *annoyed* could be linked to the rising unemployment rate and death toll. Results for other languages are provided in Appendix B.

2) Sentiments Variation of Different Countries Over Days. We chose the USA as an example to illustrate how sentiments vary over days in Fig. 3.

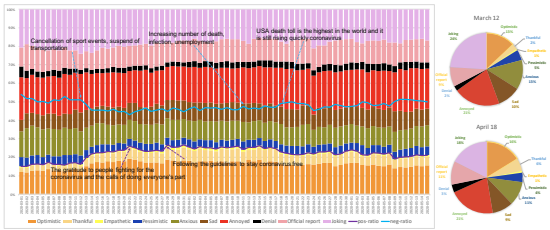


Figure 3: Sentiment variation in USA over time. Each bar shows the distribution of sentiments on one day (Better zoom in the spikes).

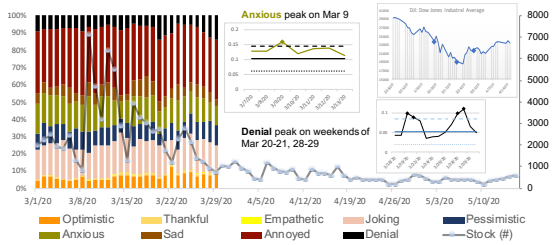
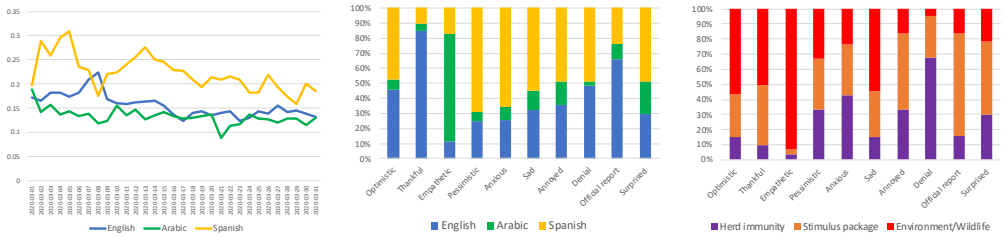


Figure 4: Sentiments variation on the stock market. We show the sentiment results when the topics were intensively discussed (around the peak of the volume curve in the background.

The blue and purple curves represent positive (sum of *optimistic, thankful, empathetic* in yellow at different intensities) and negative (sum of *pessimistic, anxious, sad, annoyed, denial* in blue at different intensities) sentiments, respectively. We observed that the proportion of negative emotions was consistently higher than that of positive emotions. On March 12, people expressed *annoyed* and *anxious* sentiments (see the pie charts) as normal life was affected by the coronavirus, including the cancellation of sports events and suspension of transportation. On March 21, positive emotions slightly increased as people expressed gratitude for the efforts of healthcare workers. However, negative emotions rose again due to increasing rates of death, infection, and unemployment on April 11. Results for other countries are provided in Appendix B.

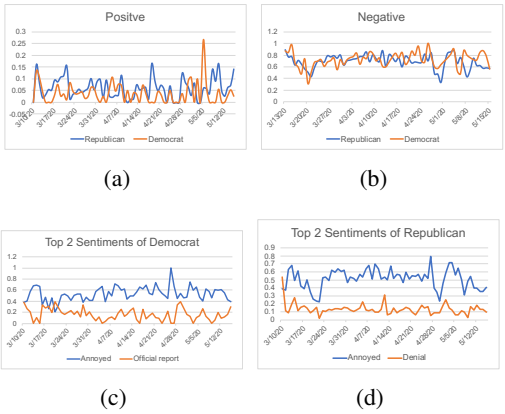
3) Sentiments Variation of Topics Over Days. We analyzed the sentiment regarding the topic *stock market* in Fig. 4. It collapsed on March 9 when the peak of the discussion was reached. *Anxious* reached a high value, surpassing the mean+2*std (above the black dashed line, where the black line represents the mean, and the dotted line is the mean-2*std). On March 12, the DJI (Dow Jones Index) experienced its worst day since 1987, plunging about 10% (triggering the second time breakers). On the weekends of March 20-21 and March 28-29, the spikes of *denial* were higher than the blue dashed line, reflecting the collapse of the stock market. Results for more topics, such as herd immunity, economic stimulus are discussed in Appendix B.

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



(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 two political parties. (a) and (b) are the trend of positive and negative sentiment. (c) and (d) show the top two sentiments over time for political parties.

and Arabic, which was possibly related to cultures and religions. Fig. 5 (b) indicated that *joking* was 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-occured with *denial*, while in the stimulus package, jokes were made with *official* reports. When discussing the environment, *joking* and *empathetic* co-occured significantly.

5) Analyzing the Public's Attitude towards Two Political Parties. In Fig. 6 (a) and (b), we depicted the trends in positive and negative sentiments for two political parties in the U.S. By analyzing tweets, we observed that the Democratic party expressed support for multiple rounds of economic stimulus, increased government spending, investment, expanded unemployment, and health insurance. On the other hand, the Republican party favored tax cuts and subsidies for 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 largely due to the postponement or denial of coronavirus relief measures. Similarly, denial sentiment reached its peak on March 10, 2020, arising from conflicts between the previous president and Democrats regarding a stimulus package. The Democrat party experienced a spike in annoyance sentiment on April 26, 2020, which could be linked to the GOP's insertion of $174 billion in tax breaks favoring the wealthy.

In summary, our analysis of sentiment variation across different languages, countries, COVID-19-related topics, and political parties provided valuable insights. We explored how diverse linguistic backgrounds influence emotional expressions, identified regional sentiment trends for tailored responses, unraveled emotional dynamics around pandemic-related topics, and tracked evolving sentiments toward political parties. These findings contributed to a comprehensive understanding of public reactions, aiding informed decision-making for governments, healthcare organizations, and policymakers during the global health crisis.

6 Conclusion

This paper introduces SenWave, a comprehensive benchmark dataset for fine-grained sentiment analysis of COVID-19 tweets. The contributions include a large annotated dataset comprising 20,000 labeled English and Arabic tweets with 10 fine-grained categories, along with 105 million unlabeled COVID-19 tweets in five languages. The study utilizes Transformer-based models as multi-label classifiers, providing detailed analyses and revealing insights into the evolving emotional landscape across different languages, countries, and topics. We employ ChatGPT to demonstrate the dataset's availability in zero- and few-shot settings. SenWave stands as a valuable resource for diverse sentiment analysis tasks requiring fine-grained emotions.

7 Limitations, Ethics & Potentiality

In this section, we introduce the limitations and ethics of our dataset, followed by a discussion of potentiality within NLP community.

Limitations. While SenWave provides a substantial collection of tweets (105 million), it is comparatively smaller than the BillionCOV dataset (Lamsal et al., 2023), which comprises over a billion COVID-19 tweets and was used for efficient hydration. Our sentiment analysis focuses on the outbreak period, and we defer exploration of post-COVID sentiment for future research. Although we gathered tweets in the top five languages, sentiments from other languages or specific regions may not be adequately represented. Additionally, the use of Twitter's API for data collection might introduce biases, as the tweets may not precisely reflect sentiments across the entire population.

Ethics. When conducting sentiment analysis on social media data, ethical considerations such as privacy, consent, and data protection are paramount. To ensure compliance with Twitter's Terms of Service and FAIR principles, any user-relevant information is removed. The dataset is licensed under Apache-2.0 license, which allows for the sharing and adaptation of the dataset under certain conditions. It is essential to acknowledge that tweets can mirror societal biases, encompassing factors like gender, race, and socioeconomic status, which may not be explicitly addressed during data collection and analysis. For example, in our analysis of public sentiments towards political parties, we refrain from inferring users' political leanings but focus on analyzing sentiments related to political parties concerning COVID-19 actions, such as stimulus packages, government spending, investment, unemployment, and health insurance. Our dataset is intended for research purposes only.

Potentiality. The SenWave dataset is poised to advance fine-grained sentiment analysis on intricate events within the NLP community. The extensive analysis of a vast pool of unlabeled data presents valuable insights for policymakers, healthcare organizations, and researchers, enabling them to make informed decisions, implement targeted interventions, and address public concerns effectively during global health crises. Moreover, given the imbalanced nature of labels in our dataset, it serves as a valuable resource for tackling the label imbalance problem in multi-label classification tasks on the SenWave dataset.

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 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*.

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*.

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*.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196.

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*.

Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shuang Li, Philip S. Yu, and Lijie Wen. 2023. Amr-based network for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 322–337.

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)*.

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*.

Zheng Zhang, Zili Zhou, and Yanna Wang. 2022. Ssegcn: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4925.

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*.

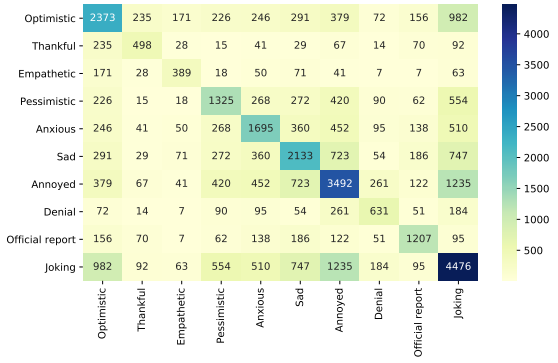
A Appendix: Dataset Information

1.1 Data Annotation

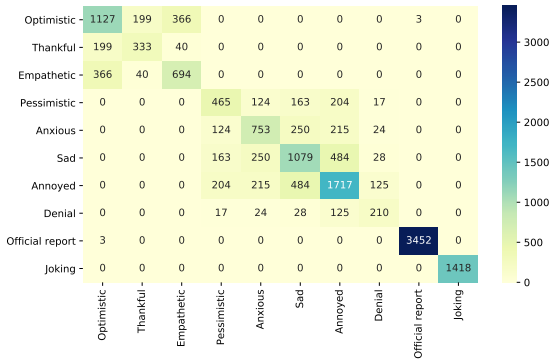
Our data was labeled by Lucidya which is an AI-based company with rich experience in organizing data annotation projects. The annotators were the native speakers or the fluent speakers. Each annotator was trained with example tweets with suggested categories to guide the annotation process. Each tweet was in dependently labeled by at least three annotators and paid 0.6 US dollars.

Table 7: 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.

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.

1.2 Label Co-occurrence Relations of English and Arabic Data

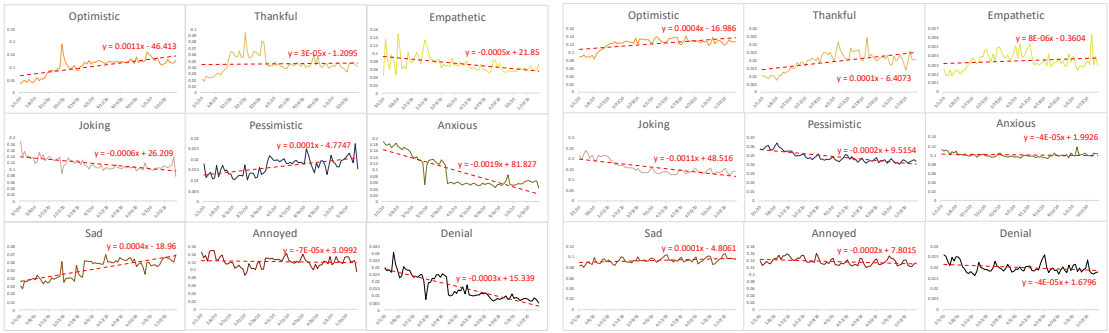
We utilized label co-occurrence heatmaps to illustrate the interrelationships between sentiment labels in both the English and Arabic datasets. In Fig. 7 (a), the complexity of label co-occurrence in the English dataset was evident, underscoring the intricacy of multi-label classification challenges. On the other hand, Fig. 7 (b) revealed that the sentiment *Official* predominates in the Arabic dataset, reflecting the substantial influence of decisions made by the Saudi government. This disparity in label co-occurrence patterns highlighted the nuanced nature of sentiment expression across different languages and cultural contexts.

1.3 Label Distribution Variance

The label distribution variation in labeled data can be attributed to distinct cultural backgrounds. The prevalence of the *joking* label was higher in English tweets compared to Arabic, while the *empathetic* label exhibited the opposite trend. Conversely, predictions on unlabeled data indicated a similar trend among English, Arabic, and Spanish, where Spanish had the highest prevalence, followed by English and Arabic.

1.4 Data Maintenance

To ensure the longevity and usability of our data repositories, we are committed to implementing a robust maintenance plan for our GitHub repository. This plan will encompass regular updates to keep



(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.

the repository current, addressing reported issues promptly, and implementing measures to enhance overall usability. Our goal is to provide a reliable and accessible resource for the benefit of future users and researchers in the field.

1.5 New Label: Official Report

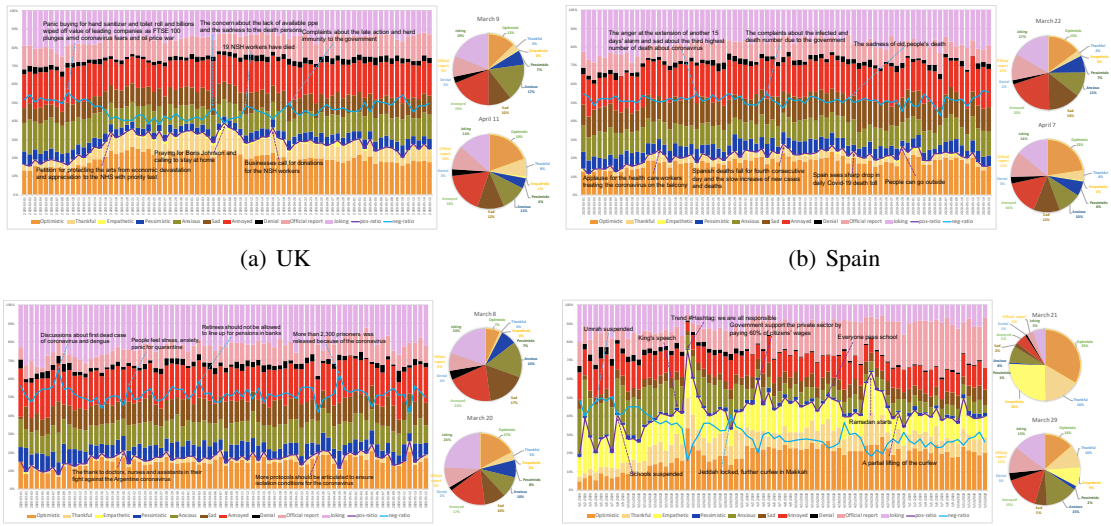
The choice of the "official report" category as a separate category stems from the fact that governments or official organizations, such as WHO, often release factual information about COVID-19, including confirmed cases, deaths, vaccine doses administered, and epidemic prevention policies. These types of tweets do not fit neatly into the "positive," "negative," or "neutral" labels; rather, they represent objective reporting of facts.

1.6 Dataset Comparison with Existing Works

The literature of sentiemnt analysis can be categorized into coarse-grained (e.g., positive, negative, neutral) and fine-grained emotions (e.g., optimistic, pessimistic, anxious, fear, joy, happiness) domains. Additionally, there's a dataset, GoEmotions, with 27 labels. A more in-depth comparison is outlined below: 1) Complexity of COVID-19 Events. The COVID-19 pandemic is a multifaceted public health event, demanding a nuanced understanding of public sentiment. Labels like positive, neutral, and negative are insufficient for capturing the complexity. 2) Suitability for COVID-19 Context. Existing labels from literature, such as joy, love, trust, amusement, pride, joy, and love, may not be suitable for COVID-19-related discussions. 3) Specialized Labels. Labels like hate and counter-hate are specialized for anti-Asian hate and counter-hate detection during the COVID-19 crisis. Similarly, COVID-19 events exhibit unique characteristics, including a significant number of official reports released by governments as well as ironic or ridiculous sentiments.

1.7 Dataset Reliability Evaluation With ChatGPT

In our multi-label text classification experiments using ChatGPT-3.5, we conducted zero-shot and few-shot classification. In the zero-shot setting, no labeled tweets were provided to ChatGPT, only the prompt and label-removed data were used. For the few-shot classification, a minimal set of labeled tweets (38 out of 10,000) the prompt and label-



(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)

removed data were provided to ChatGPT. The 38 tweets were randomly selected to ensure coverage of all labels. The designed prompts used in these experiments are detailed in Table 7.

B Appendix: More Interesting Findings

We first introduced the data processing of the used unlabeled tweets. We then presented more analyzed results about sentiment variation on the unlabed data including: 1) **how sentiment varied in different languages**; 2) **how sentiment varied in different countries**; and 3) **how sentiment varied in different topics**.

2.1 Unlabeled Data Processing for Sentiment Analysis

For the unlabeled data, we first select the relevant tweets towards the specific targets based on the properties of data, or the pre-defined keywords. Secondly, we use the well-trained classifiers to predict the labels of selected tweets. Lastly, we analyze the predicted results from the mentioned aspects above and draw the corresponding figures.

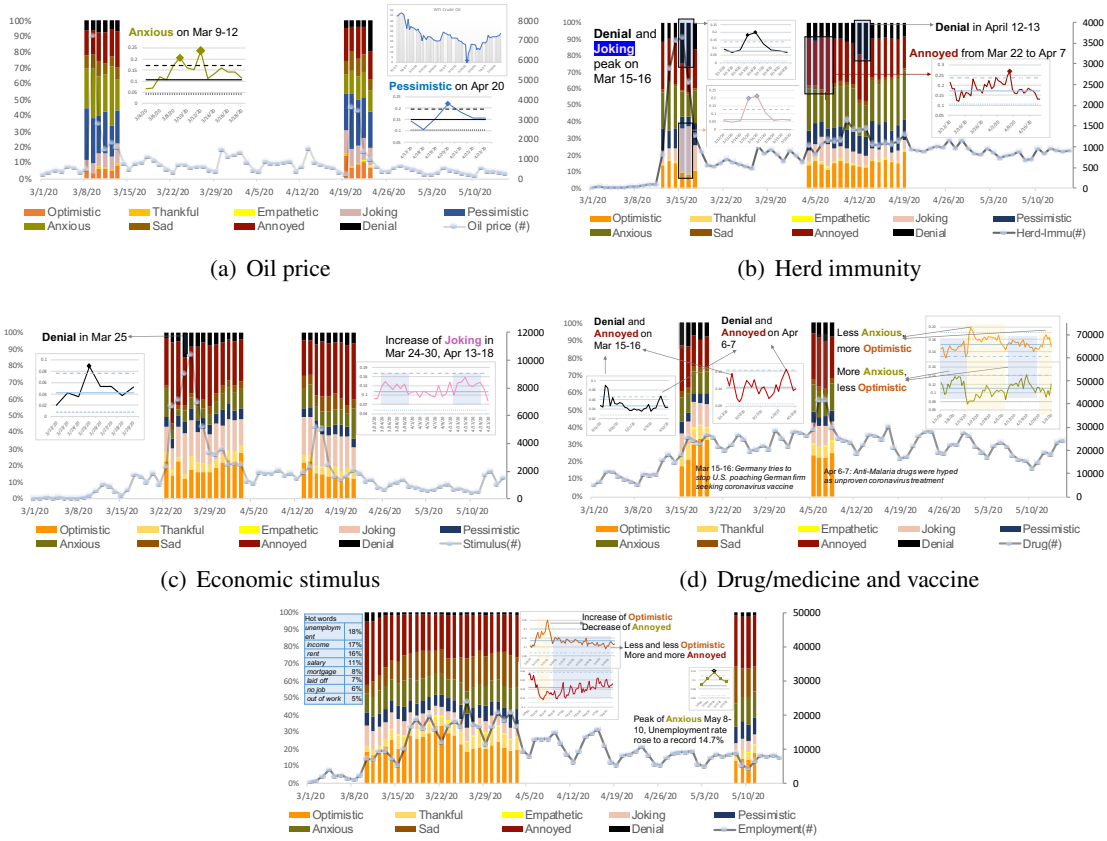
2.2 Sentiment Variation of Different Languages Over Days

The results of Arabic tweets shown in Fig. 8 (a) demonstrated significant variations in all categories of emotions. In particular, *optimistic* kept rising up, and *anxious, denial* and *joking* were falling down. The *sad* emotion kept 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* was observed in French tweets, as shown in Fig. 8 (c). However, the other emotions became stable, except for the decline of *joking* and the sudden increase of *denial*, attributed to the conspiracy theory about the lab source of coronavirus. Italian tweets also showed a weak increase or decrease trend in most of the emotions, as shown in Fig. 8 (d), except for those in *thankful* and *empathetic*.

2.3 Sentiments Variation of Different Countries Over Days

Fig. 9 (a) illustrated the sentiment dynamics in the UK. On March 9, negative emotions surged due to panic buying of essential items and concerns about the coronavirus and the oil price war, causing a decline in the FTSE 100. Following the implementation of various coronavirus measures, positive sentiments experienced a significant rise.

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



(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).

pie chart).

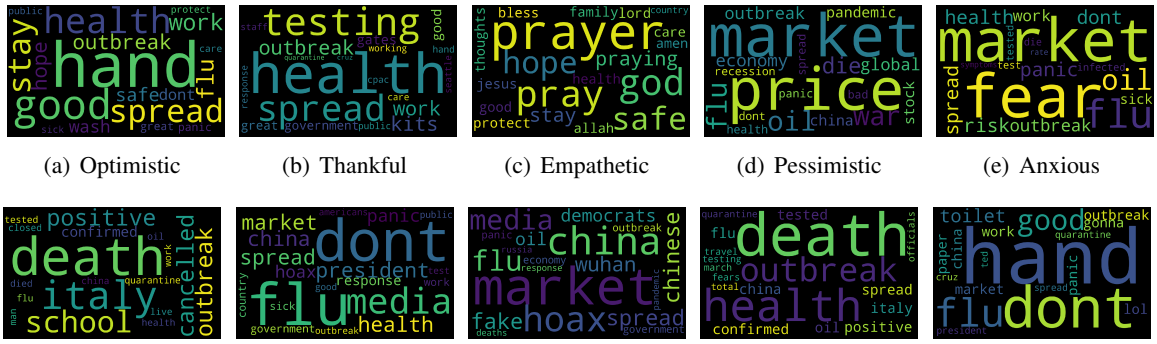
In Argentina (Fig. 9 (c)), the proportion of negative emotions was very close to 0.5 even much higher on some days. On March 8, the discussions about the first death case of coronavirus and dengue were focused on leading to the increase of *anxious*, *sad*, and *annoyed* (see pie chart at the right-hand). 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) illustrated a notably stronger positive sentiment in Saudi Arabia compared to other countries or regions. Particularly, from March 13 onward, there was a surge in positive emotions coinciding with numerous decisions made by the Saudi government. The peak was observed on March 21, in response 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 occurred on April 23-24, coinciding with the start of Ramadan.

2.4 Sentiments Variation of Studied Topics Over Days

As depicted in Fig. 10 (a), the discussion on oil prices peaked on March 9. The sharp decline in crude oil prices led to a substantial increase in *anxious* sentiments from March 9 to 12. However, this period did not mark the peak of anxiety. On April 21, when the crude oil price hit an 18-year low, highlighted on the WTI crude oil curve, discussions were particularly dominated by *pessimistic* sentiments.

Fig. 10 (b) highlighted the topic of herd immunity, which rapidly gained traction on March 14-15 following the UK government's initial consideration on March 13. During the intensive discussions from March 13 to 17, *denial* and *joking* were significantly observed on March 15-16. The discussion continued with a notable increase in *annoyed* from March 22 to April 7, causing another rise in *denial* on April 12-13.



(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

As illustrated in Fig. 10 (c), the topic of economic stimulus reached its peak on March 26 when the US Senate passed a historic $2tn relief package, with another peak on April 15-16 when the checks were received. Surprisingly, during the discussion on March 23-26, positivity was lower compared to other days, and *denial* was significant on March 25. Many tweets under this topic expressed sentiments such as "This is not enough", "US economy is tanking", and "The pandemic is getting worse". Examining *joking*, increases were observed on March 24-30 and April 13-18.

Fig. 10 (d) demonstrated that the topic of drug/medicine/vaccine generated the largest amount of discussion among the five topics, reaching 20-40K in daily volume. This topic gained prominence due to the global outbreak around March 10. Two events caused significant *denial* and *annoyed*: first, on March 15-16, when Germany tried to stop the U.S. from poaching German firms seeking coronavirus vaccines, and second, on April 6-7, when Anti-Malaria drugs were hyped as an unproven coronavirus treatment.

In Fig. 10 (e), the topic of employment/job covered keywords such as unemployment, income, rent, salary, mortgage, laid off, no job/work, etc. In March, there was an increase in *optimistic* and a decrease in *annoyed*. However, in April-May, there was less *optimistic* sentiment and an increase in *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.

2.5 Meanings of Analyzed Results

The analysis on the unlabeled tweets can identify patterns of fear or anxiety to reduce the anxiety, monitor public sentiments toward government actions and policies to improve satisfaction, and detect emerging concerns or issues to take the proper actions quickly. Particularly we can know 1) how people in different linguistic backgrounds express their emotions; 2) identify regional sentiment trends beneficial for governments, healthcare organizations, and businesses; 3) how sentiments differ across various topics can provide insights into which aspects of the pandemic were polarizing or emotionally charged; 4) gauge public opinion and track how political responses to the pandemic

influence public sentiment.

C Appendix: Hot Words Visualization

We presented 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* was 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* was shown with "pray", "hope", "god", and "safe". The class *pessimistic* was reflected in the economy market, oil market, and a large number of deaths. These hot words were 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* was displayed with "dont" and "flu" while the class *denial* was 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* was represented with الوقاية (protection), نمنع (prevent), علاج (treatment), and الخير (the good). The class *thankful* showed شكرا (Thanks), السعودية(Saudi), سلمان (Salman, the king), and الكـ ويت(Kuwait), which reflected how people were happy with governments actions against Covid-19. The *empathetic* words showed the prayers to Allah for protecting the people and countries. The class *pessimistic* represented الناس (people), الشعب(commune), الحجر (quarantine), and أزمة(crisis). In *anxious* class, the words انتشار (spread), نخاف (fear), استغفار(asking forgiveness) were the popular words. The class *annoyed* represented المرض (disease), الصين(China), ايران (Iran), where the first case appeared in Saudi came from Iran. العالم (The world), حرب (war), مؤامرة (conspiracy) were the hot words in *denial* class which reflected how people think about this virus. The words in *joking* were الحجر(quarantine), البيت (house), الناس (people), and ابريل(April).