Emotion analysis and detection during COVID-19

Anonymous ACL submission

Abstract

Understanding emotions that people express during large-scale crises helps inform policy makers and first responders about the emotional states of the population as well as provide emotional support to those who need such support. We present CovidEmo, a dataset of ∼3,000 English tweets labeled with emotions and temporally distributed across 18 months. Our analyses reveal the emotional toll caused by COVID-19, and changes of the social narrative and associated emotions over time. Motivated by the time-sensitive nature of crises and the cost of large-scale annotation efforts, we examine how well large pre-trained language models generalize across domains and timeline in the task of perceived emotion prediction in the context of COVID-19. Our analyses suggest that cross-domain information transfers occur, yet there are still significant gaps. We propose semi-supervised learning as a way to bridge this gap, obtaining significantly better performance using unlabeled data from the target domain.

1 Introduction

We live in unprecedented times caused by a coronavirus: the COVID-19 pandemic. This pandemic has forced extremely rapid changes in our daily lives in the push to stem the spread of the COVID-19 virus. Many of us have been uprooted, disrupted and distanced from family, friends and colleagues. We have transitioned in no time into a world that is suddenly more virtual than personal, sacrificing many of the daily rhythms and joys of life. These events coupled with the dramatic lifestyle changes consequently led to vast amounts of data generated on social media platforms such as Twitter. Understanding emotions that people increasingly express on social media during large-scale crises can have wide-ranging implications, from promoting a deeper understanding of the society to informing policy makers and first responders about the emotional states of the population (Dennis et al., 2006; Fraustino et al., 2012). In Natural Language Processing, multiple datasets have been proposed to detect emotions on social media (Mohammad, 2012; Wang et al., 2012; Mohammad and Kiritchenko, 2015; Volkova and Bachrach, 2016; Abdul-Mageed and Ungar, 2017; Demszky et al., 2020), including from hurricane disasters (Schulz et al., 2013; Desai et al., 2020). Recent studies propose the investigation of emotions in COVID-19 (Kabir and Madria, 2021; Imran et al., 2020; Ashokkumar and Pennebaker, 2021; Ng et al., 2020). Ashokkumar and Pennebaker (2021) explore the expression of emotions to analyze the psychological shifts caused by COVID. Nevertheless, their assessment of emotions is based on lexical cues and does not capture implicit emotions. In contrast, Ng et al. (2020) train deep learning emotion classifiers outside the COVID-19 and apply it on COVID data to analyze emotions. However, our analyses suggest that emotion detection models trained outside the COVID-19 domain struggle to transfer information.

In this paper, we explore the detection of perceived fine-grained emotions during the COVID-19 pandemic to answer two research questions. First, from a **social** point of view, each crisis is situated in its own unique social context (Palen and Anderson, 2016), triggering distinct emotions, and impacting different populations in vastly distinct ways. COVID-19 is a crisis that has dominated the world stage and influenced every aspect of human life. *What are the emotions expressed through social media, and how do they change over time*? Second, from a **system** point of view, modern data-driven emotion prediction systems are trained on large, annotated datasets. *How well can models learn from existing resources since timely annotation of fine-grained emotions can be costly to accumulate as new crises arise, and how well do models generalize as a crisis unfolds through different stages?*

To answer these questions, we introduce

<USER> Please resign, you are the master of misleading who started politicizing the public health crisis. You are a part of the problems the world is facing!	anger, disgust, sadness
'Perfect storm': Haiti COVID-19 peak set to collide with hurricanes. <URL>	fear, sadness
The German government is taking all kind of measures to protect its people while the Dutch government does not care about their people #corona	surprise, trust, anger

Table 1: Examples from CovidEmo annotated with the Plutchik-8 emotions.

CovidEmo, a dataset of ~3K tweets in English annotated with Plutchik-8 emotions (Plutchik, 2001); examples are shown in Table 1. Our dataset provides an ideal test bed to examine how well modern NLP models generalize across domains and crises in the task of perceived emotion prediction. Moreover, CovidEmo is temporally distributed across 18 months, which enables the exploration of distributional shifts that occurred from the start of the pandemic. Our analysis reveals that the co-occurrence and distribution of emotions are drastically different from natural disasters such as hurricanes (Desai et al., 2020). However, while Desai et al. (2020) pointed out that emotion distributions are fairly consistent across hurricanes, in CovidEmo we observe a different phenomenon: as COVID-19 progressed, we note considerable distributional shifts both in the lexical and the emotion label space. Additionally, we found that politically related words are more likely to associate with negative emotions, while vaccine-related words are more likely to associate with positive ones.

We carry out a comprehensive set of experiments that evaluate model generalization under domain shift. Experimenting with large-scale pretrained language models including BERT (Devlin et al., 2019), BERTweet (Nguyen et al., 2020), and COVID-Twitter-BERT (Müller et al., 2020), we find that directly applying models trained on other emotion datasets to CovidEmo leads to poor overall performance, indicating considerable domain gaps. Our analysis also reveals two surprising findings: **1)** Performing direct transfer from a general emotion dataset such as GoEmotions (Demszky et al., 2020) attains better performance compared to transferring information from a disaster-specialized corpus such as HurricaneEmo (Desai et al., 2020), indicating the vast differences across crises. **2)** Besides the inter-domain gaps observed, we note in-domain model performance gaps along the temporal dimension as well. Specifically, we find that training a model on the first 6 months of our data and testing on the last 6 months obtains a 2% decrease in F-1 score compared to using training and testing data from the same timeframe (last six months).

Finally, we investigate methods to bridge both the inter-domain and the in-domain temporal gaps. We motivate the importance of lowering these gaps: first, due to the time-critical, dynamic nature of disasters such as COVID-19, the time needed to acquire labeled data might severely impact the early-risk assessment capabilities of the authorities and slow the relief response. Second, labeling data for every potential disaster is not feasible in terms of annotation costs. To this end, we leverage Noisy Student Training (Xie et al., 2020), a semi-supervised learning technique utilizing the readily available COVID-19 unlabeled data, and the non-COVID labeled data, to obtain a better emotion detection model. This improves the performance of the vanilla models significantly, by 1.5% on average.

We summarize our contributions as follows: **1)** We introduce CovidEmo, an emotion corpus containing ∼3K tweets streamed during the COVID-19 pandemic, which enables the exploration of model generalization across domains, as well as between different time periods of the same domain. **2)** We perform a comprehensive analysis of emotion expression in CovidEmo, indicating various particularities and comparing our corpus with other datasets in the literature. **3)** We observe considerable domain gaps and offer potential explanations into why models struggle to transfer information. **4)** We bridge these gaps using semi-supervised learning. We will release our data and models upon publication.

2 Data

2.1 Data collection

Preprocessing. We sample $129,820$ English tweets from Chen et al. (2020)'s ongoing collection of tweets related to the COVID-19 pandemic, starting from January 2020 until June 2021[1]. Our sampling strategy involves selecting an equal number of tweets each month in the time period mentioned above. The tweets are anonymized by re-

[1]We use the Twarc software to obtain the tweet texts, and FastText (Joulin et al., 2017) for language identification.

Emotion	Content words/Hashtags
disgust	**Content words:** disgusting, fucking, million, trump, dead, shit, president, america, china, done **Hashtags:** #hongkong, #gop, #factsmatter, #ccp, #china, #wuhan, #covid19
anger	**Content words:** fuck, evil, bullshit, stupid, idiot, damn, obama, church, lying **Hashtags:** #marr, #covidiots, #trumpvirus, #torycorruption, #skynews, #qanon, #nh, #jacksonville, #gop, #factsmatter
fear	**Content words:** scared, exam, dangerous, infected, confirmed, worse, sir, wuhan, risk, rate **Hashtags:** #stopcovidlies, #jeeneet, #antistudentmodigovt, #health, #wuhan, #china, #stayhome, #covid19
sadness	**Content words:** sad, cry, died, suffering, toll, record, sorry, feel, tested, facing **Hashtags:** #notmychild, #quarantine, #rip, #pregnant, #italy, #healthcare, #freepalestine, #askktr, #wuhan, #vaccine
anticipation	**Content words:** effort, christmas, available, join, start, future, vaccination, vaccinated, coming, open **Hashtags:** #stayhomestaysafe, #pregnant, #postponeinicet, #nyc, #launchzone, #fred2020, #cow, #what-shappeninginmyanmar, #ethereum, #bcpoli
trust	**Content words:** working, support, safe, help, say, being, world, vaccine, good, more **Hashtags:** #stayhome, #staysafe, #covid19, #lockdown, #china
joy	**Content words:** grateful, beautiful, thanks, happy, love, great, little, morning, good **Hashtags:** #taiwan, #innovation, #breaking, #staysafe, #stayathome, #stayhome, #wearamask, #lockdown, #covid19
surprise	**Content words:** believe, year, lockdown, new, china, virus, day, america, covid19, get **Hashtags:** #china, #covid19

Table 2: Content words and hashtags most associated with each Plutchik-8 emotion.

placing twitter usernames with <USER> and links with <URL>, following Cachola et al. (2018). Additionally, prior work found that even in disaster contexts, the fraction of tweets expressing an emotion is small (Desai et al., 2020), thus annotating randomly sampled tweets would be costly and unproductive. Therefore, we follow their work to obtain tweets that are more likely to contain emotions for annotation. Concretely, we ensure that each tweet encompasses at least one word from EmoLex (Mohammad and Turney, 2013), a lexicon of ~10K words in various languages annotated with emotion labels. After this filtering process, we obtain 89,274 tweets. As stated in Desai et al. (2020), this filtering is soft, i.e., does not filter out tweets with weak or implicit emotions.

Annotation and quality control. We randomly sample 5,500 tweets from this data and use Amazon Mechanical Turk to crowdsource Plutchik-8 emotions: *anger, anticipation, joy, trust, fear, surprise, sadness, disgust*. We allow multiple selection, as well as a *none of the above* option in case no emotion is perceived. During the annotation process, we determine the inter-annotator agreement using the Plutchik Emotion Agreement (PEA) metric that take into account emotion proximity on the Plutchik wheel (Desai et al., 2020).

We use a qualification process for quality control and training. Specifically, two members of our research team annotated a small set of tweets, from which we selected 20 examples where both anno-

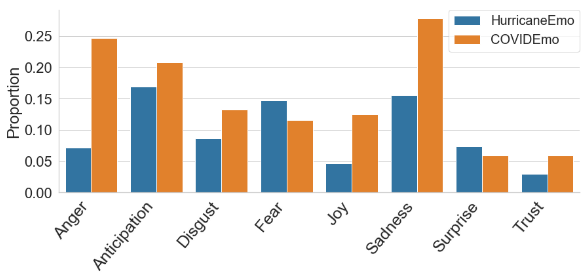

Figure 1: Emotion distribution of two types of crises: hurricanes and the COVID-19 pandemic.

tators agree on the emotions. We qualify workers whose annotations attain high agreement with ours (PEA>75.00) calculated against our annotations. This results in a highly capable pool of workers for the main task. Additionally, we exclude annotations from workers who have very poor agreement with others (Cachola et al., 2018; Desai et al., 2020) (those whose PEA scores are below the 80th percentile compared to others). Each tweet has at least 2 annotations after filtering.

We aggregate labels such that an emotion is considered present if *at least two workers* perceived the emotion. This resulted in 2,847 tweets in CovidEmo with an average, per-worker PEA score of 84.05, indicating high inter-annotator agreement.

2.2 Analysis

Emotion distribution. We show the general distribution of Plutchik-8 emotions in CovidEmo in Figure 1. We note that the percentage of neg-

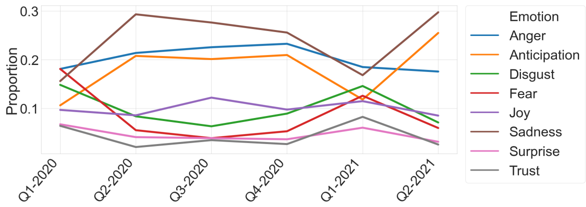

Figure 2: Emotion distribution in COVIDEMO over time (by quarter).

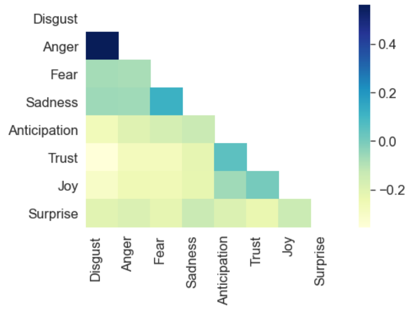

Figure 3: Emotion co-occurrence in COVIDEMO.

ative emotions (*disgust, anger, fear, sadness*) is much higher than that of positive emotions (*trust, joy*), consistent of the emotional toll of COVID-19. Next, we draw comparisons between the emotion distribution in COVID-19 and that of natural disasters, specifically HurricaneEmo (Desai et al., 2020), shown in Figure 1. We make a few observations: First, the tweets in COVIDEMO contain a higher emotion proportion across six out of the eight total emotions, indicating that COVID-19 prompted an increased multi-label emotional response compared to natural disasters. Second, the sadness emotion is almost twice more represented in COVIDEMO compared to HurricaneEmo, whereas we see as much as a four-fold increase in the representation of anger. Finally, we observe that *anticipation* is much more prevalent in HurricaneEmo and a lot less frequent in the pandemic, which matched the COVID-19 reality that it is hard to anticipate events/facts.

We also show **emotion distribution across time** in Figure 2, obtained grouping the tweets by quarter (e.g., Q1-2020 encompasses the first three months of 2020). We observe that the label distribution varies significantly from quarter to quarter, denoting potential changes in the discussion topics or the overall feelings of the masses. Notably, we note proportion variations as high as 12% in consecutive quarters. For instance, the proportion of the sadness emotion increases by as much as 12% in the second quarter of 2020 compared to the first quarter. Moreover, we see the opposite trend in the fear emotion, whose proportion decreases by 10% percent. One potential explanation could be that the first shock that COVID-19 produced enacted fear into people (Q1 2020). However, as people started to get accustomed to the lockdown, the fear slowly turned into sadness.

Emotion co-occurrence. Figure 3 depicts how emotions co-occur with one another in COVIDEMO. For each emotion pair, we compute the Pearson correlation coefficient. Overall, we observe stronger correlation between emotions in the same positive/negative categories. For example, (*anger, disgust*) and (*sadness, fear*) appear much more frequently than (*anger, anticipation*) and (*anger, joy*). Table 1 shows samples from COVIDEMO with multiple emotions perceived. Notably, in many cases lexical cues alone cannot account for the emotions, as evident in the second example. Although the word "perfect" suggest optimism, the annotations are nowhere near positive. In the third tweet, there is a co-occurrence of polarizing emotion because the tweet deals with a positive and a negative situation at the same time.

	ANG	ANT	DIS	FEA	JOY	SAD	SUR	TRU
DEV	327	296	163	179	186	388	86	89
TEST	374	296	214	149	170	403	83	78

Table 3: Validation and test set splits for eight Plutchik-8 emotions, including including anger (ang), anticipation (ant), disapproval (dis), fear (fea), joy, sadness (sad), surprise (sur), trust (tru).

Lexical analysis. To understand better what topics or events are associated with each emotion, we perform a lexical analysis to examine the co-occurrence between content words (nouns, verbs, adjectives and adverbs), hashtags and perceived emotions. In particular, we calculate the log odds ratios ($\log(P(w|e)/P(w|\neg e))$ (Nye and Nenkova, 2015) with a frequency threshold of 10 for lemmatized content words and 2 for hashtags. Table 2 shows the highest ranked content words and hashtags for each emotion category. We notice that politically or country-oriented words are more likely to associate with negative emotions *(president, america, china)*, while vaccine-related words are more likely to associate with positive emotions.

2.3 Benchmark Dataset

To enhance reproducibility and aid the progress on understanding the expression of emotion in the

COVID-19 context, we use CovidEmo as a benchmark dataset for perceived emotions. We split our data into a development and testing split, as shown in Table 3. We also note that the data is evenly distributed across the time axis, with an equal number of 158 tweets for each of the 18 months that our dataset spans. As mentioned previously, disasters are time-critical events, and since our goal is to examine the emergence of such disasters, we mainly focus on domain adaptation techniques, hence we omit creating a training set.

3 Domain Transfer Assessment

Using CovidEmo, we evaluate the ability of modern NLP models to transfer information from existing sources with annotated emotions in an inter-domain setting for perceived emotion detection, and if models generalize temporally in the same larger context (in-domain temporal transfer).

3.1 Our Framework

We consider a dataset $\mathcal{S}$ labeled with emotions, and another collection of labeled examples $\mathcal{T}$ from a different domain. We aim to assess how well large pre-trained language models can transfer information from the domain of $\mathcal{S}$ to the domain of $\mathcal{T}$. To this end, we train our models on $\mathcal{S}$, then evaluate the performance on the test set of $\mathcal{T}$. In our framework, $\mathcal{T}$ is CovidEmo for the inter-domain experiments, or a temporal slice of CovidEmo for the temporal experiments. Due to the uneven label distribution and the multi-label nature of the data, we develop binary classifiers for each emotion following Desai et al. (2020).

Methods. Motivated by the tremendous success of large pre-trained masked language models, we use the following models: **1)** BERT (Devlin et al., 2019) base uncased model trained on Wikipedia and BookCorpus (Zhu et al., 2015), **2)** BertTweet (Nguyen et al., 2020) model trained on 850M english tweets, and **3)** COVID-Twitter-BERT (CT-BERT) (Müller et al., 2020) trained on 97M tweets. Additionally, we also employ a basic lexicon-based classification approach, **4)** EmoLex (Mohammad and Turney, 2013) is the word-associated lexicon mentioned previously in the paper. In this approach, if a tweet contains a word annotated with an emotion e in EmoLex, then we assign e as a label for the tweet.

Experimental setup. We perform all our experiments on an Nvidia P100 GPU. To report the performance, we average the F-1s of 5 different runs and report the average value. We present in Appendix A detailed information about the hyperparameters used for the best models. Additionally, in Appendix B, we indicate the hyperparameter search space explored, as well as model running times.

3.2 Inter-domain Transfer

Our first domain transfer assessment explores how well emotion detection models trained outside our domain generalize to the COVID context. We consider two well-established datasets for training. First, we experiment with GoEmotions (Demszky et al., 2020), a dataset from the general Reddit domain annotated with 28 emotions and the neutral class. The emotion space in GoEmotions differs slightly from our Plutchik-8 setup, hence we perform a mapping[2] between the emotions in GoEmotions and the Plutchik-8 emotions. Second, we use HurricaneEmo (Desai et al., 2020), a Twitter dataset collected from natural disasters such as hurricanes and labeled with fine-grained emotions. HurricaneEmo provides Plutchik-8 labels.

Results. We show the results obtained in Table 4. Here, we denote by *M-DS* the model *M* trained on dataset *DS* and tested on CovidEmo. We emphasize a surprising finding: **models trained on a general domain (GoEmotions) generalize better on CovidEmo compared to models trained on natural disasters such as hurricanes (HurricaneEmo).** In fact, the performance gaps between GoEmotions and HurricaneEmo are vast, and we see as much as 0.20 differences in average macro F-1. At the same time, we note that our basic lexicon-based Emolex approach outperforms the HurricaneEmo transfer models. This result hints to a sizeable divergence between crises such as hurricanes and COVID-19. The CT-BERT model improves the performance by 1% on average (with statistical significance), compared to BERTweet which only obtained marginal improvements. Although both are trained on Twitter data, we postulate that CT-BERT likely benefited from COVID-related biases that the model manages to leverage.

[2] GoEmotions Mapping: Anger → Anger, Disgust → Disgust, Joy → Joy, Sadness → Sadness, Fear → Fear, Nervousness, Desire → Anticipation, Surprise → Surprise, Admiration → Trust.

MODEL	ANG	ANT	DIS	FEA	JOY	SAD	SUR	TRU	AVG
BERT-GOEMOTIONS	0.735	0.589	0.624	0.625	0.722	0.687	0.588	0.540	0.635
BERT-HURRICANEEMO	0.592	0.339	0.563	0.398	0.385	0.467	0.403	0.347	0.433
BERTWEET-GOEMOTIONS	0.752	0.534	0.631	0.629	0.709	0.708	0.624	0.537	0.637
BERTWEET-HURRICANEEMO	0.677	0.346	0.540	0.311	0.299	0.494	0.354	0.418	0.435
CTBERT-GOEMOTIONS	0.735	0.577	0.629	0.644	0.725	0.717	0.617	0.520	0.644†
CTBERT-HURRICANEEMO	0.655	0.366	0.471	0.311	0.341	0.447	0.243	0.349	0.406
EMOLEX	0.57	0.517	0.547	0.551	0.543	0.560	0.458	0.414	0.504

Table 4: Direct transfer Macro F-1 scores using BERT (Devlin et al., 2019) base uncased model (BERT-*), BERTweet (Nguyen et al., 2020) (BERTWEET-*) and Covid-Twitter-BERT (CTBERT-*). The results in this table are average F-1s across 5 different runs. We assert significance† if $p < 0.05$ under a paired-t test with the vanilla BERT model.

MODEL	ANG	ANT	DIS	FEA	JOY	SAD	SUR	TRU	AVG
CTBERT-$\mathcal{F}_{tr}$	0.762	0.485	0.534	0.661	0.705	0.673	0.492	0.492	0.600
CTBERT-$\mathcal{L}_{tr}$	0.769	0.631	0.498	0.668	0.781	0.724	0.493	0.502	0.633†

Table 5: Macro F-1 scores using in-domain temporal adaptation. The CTBERT-$\mathcal{L}_{tr}$ improvements are statistically significant†.

	Cosine Similarity			Jensen-Shannon Divergence		
	COVID	GOEMOTIONS	HURRICANEEMO	COVID	GOEMOTIONS	HURRICANEEMO
COVID	1.0	0.346	0.243	0.0		
GOEMOTIONS		1.0	0.378	0.312	0.0	
HURRICANEEMO			1.0	0.351	0.374	0.0

Table 6: Cosine similarities and Jensen-Shannon divergence of word distributions between GoEmotions (Demszky et al., 2020), HurricaneEmo (Desai et al., 2020), and COVIDEMO.

3.3 In-domain Temporal Transfer

COVIDEMO spans a large period of time (18 months) marked by substantial narrative shifts in the society. Thus we investigate potential distributional shifts across the temporal dimension. Specifically, we aim to analyze how well models trained on past COVID-19 data generalize to a fresh batch of new data. To this end, we stage the following setup: First, we accumulate the subsets $\mathcal{F}$ and $\mathcal{L}$ corresponding to the initial six months and the last six months respectively. Denoting the development and test sets of COVIDEMO as $\mathcal{C}_{tr}$ and $\mathcal{C}_{ts}$, we create additional sets $\mathcal{L}_{tr} = \mathcal{L} \cap \mathcal{C}_{tr}$ and $\mathcal{L}_{ts} = \mathcal{L} \cap \mathcal{C}_{ts}$. Additionally, we randomly subsample $\mathcal{F}_{tr} \subset \mathcal{F}$ such that $|\mathcal{F}_{tr}| = |\mathcal{L}_{tr}|$, where $|.|$ denotes the size of a set. In this setting, we compare training on $\mathcal{F}_{tr}$ and testing on $\mathcal{L}_{ts}$ vs. training on $\mathcal{L}_{tr}$ and testing on $\mathcal{L}_{ts}$. In other words, we investigate whether model performance on COVIDEMO decreases as time passes. Here we experiment with CT-BERT (Müller et al., 2020) (since it achieved better performance in Section 3.2).

Results. Table 5 shows that **the models trained on the same time period as the testing data outperforms the model trained on a different timeframe significantly**, obtaining a Macro F-1 increase of 3.3% on average. Notably, we observe improvements as high as 7.6% in F-1 on joy and 14.6% on anticipation. Intuitively, since the model is trained on the same temporal distribution as the test set, and anticipation is closely related to ongoing events (i.e., people usually anticipate certain events), it is extremely probable that the model has been trained on similar events in the training set, so the model easily recognizes the emotion.

4 Understanding Domain Gaps

The previous section exposed significant inter-domain and temporal gaps leading to poor transfers of information between these domains. In this section, we aim to answer the following questions: Why does GoEmotions transfer better than HurricaneEmo, even though the latter is a disaster-centric dataset? How did data distribution shift during the pandemic? We hope that our insights can spur further research into bridging these gaps. In Section 5, we propose semi-supervised learning as a method to build better transfer learning models.

Inter-domain gaps. To answer the first question, we analyze the lexical differences between GoEmotions, HurricaneEmo, and COVIDEMO. In order to obtain more accurate comparisons in terms of the larger vocabulary, we use unlabeled data for HurricaneEmo and CovidEmo to match the number of examples in GoEmotions (~60K). Table 6 shows the cosine similarity and the Jensen-Shannon divergence for the frequency distribution of all content words (lower-cased and lemmatized) across the three datasets. Interestingly, the CovidEmo distribution is significantly closer to GoEmotions compared to HurricaneEmo: the cosine similarity is substantially lower (0.243 vs. 0.346) while the divergence is larger (0.312 vs. 0.351). Moreover, the HurricaneEmo distribution diverges even more from GoEmotions compared to COVID-19. These findings hint that although HurricaneEmo is closer to CovidEmo than to a general domain, the COVID-19 context is significantly more correlated with a general domain than a natural disaster one, likely due to the wide impact COVID-19 has had and a more social nature of the crisis. These findings could also explain why there are large gaps in performance between HurricaneEmo and GoEmotions transfers.

In-domain temporal gaps In Section 2.2, we revealed that the label distribution and topics discussed during COVID-19 has shifted over time. To consolidate these analyses, we carry out an additional experiment that captures distributional shifts in vocabulary. In Figure 4 we show the cosine similarities and Jensen-Shannon divergence for the frequency distributions of content words (lower-cased and lemmatized) for unlabeled tweets spanning the 18 months in our data. As time passes, we observe a constant shift in the lexical distribution of the tweets. Concretely, while the cosine similarity between the first and the second month of COVID-19 is 0.97, by the end of the 18^{th} month this value decreases significantly, getting as low as 0.63. We observe the same phenomenon in the divergence of the distributions as well. These findings emphasize the considerable temporal gaps found in long-lasting disasters such as COVID-19, and that temporal slices of the tweets can diverge significantly even though they originate from the same domain.

5 Bridging the Gaps Between Domains

As crises such as COVID-19 strike, large amounts of user-generated content are produced on social sites. However, due to the nature of disasters unfolding rapidly, the high costs needed for annotation, and the considerable distributional changes

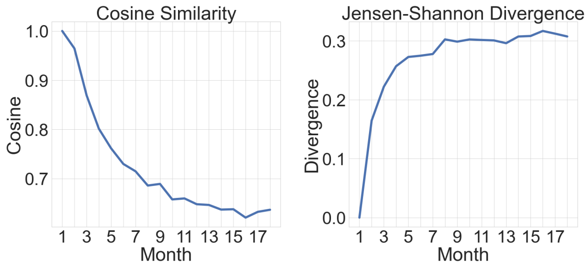

Figure 4: Cosine similarities and Jensen-Shannon divergence between the first month of COVID-19 and each subsequent month.

along the time axis, immediately obtaining labeled data from the ongoing disaster might prove infeasible. However, rapid understanding of such events is critical for rapid risk assessment and effective resource allocations. Therefore, we cannot rely on obtaining large quantities of labeled data, and we require effective domain adaptation techniques which can leverage labeled data from outside the disaster domain. However, we emphasized previously that models typically have a hard time effectively transferring information for emotion detection. We argue that even though we cannot timely obtain labels for the ongoing disaster, we can still use the large amounts of unlabeled user-generated Tweets to build better domain adaptation models. To this end, we experiment with semi-supervised learning.

Method. Noisy student training (Xie et al., 2020) is an approach leveraging knowledge distillation and self-training, which iteratively jointly trains two models in a teacher-student framework. The model leverages noised unlabeled data alongside labeled data to obtain better performance. We detail the setup we use as well as the various noising techniques in Appendix C. A vital aspect of our framework, however, is that we use unlabeled data from COVID-19. Concretely, in the inter-domain adaptation experiments, where we aim to transfer information from GoEmotions to CovidEmo, we use labeled data from GoEmotions alongside unlabeled data from CovidEmo (we make sure the model does not see any example from the test set). In the temporal setup, where we train on the first six months $\mathcal{F}_{tr}$ and test on the last six $\mathcal{F}_{ts}$, we use $\mathcal{F}_{tr}$ in conjunction with unlabeled data generated in the last six months.

Results. We show the results obtained using Noisy Student training in Table 7. **Our SSL technique bridges both the inter-domain and the**

MODEL	ANG	ANT	DIS	FEA	JOY	SAD	SUR	TRU	AVG
CTBERT-GOEMOTIONS	0.735	0.577	0.629	0.644	0.725	0.717	0.617	0.520	0.644
CTBERT-GOEMOTIONS-SSL	0.741	0.554	0.657	0.651	0.741	0.726	0.632	0.532	$0.654^{\dagger}$
CTBERT-$\mathcal{F}_{tr}$	0.762	0.485	0.534	0.661	0.705	0.673	0.492	0.492	0.600
CTBERT-$\mathcal{F}_{tr}$-SSL	0.771	0.501	0.531	0.711	0.711	0.671	0.538	0.501	$0.617^{\dagger}$
CTBERT-$\mathcal{L}_{tr}$	0.769	0.631	0.498	0.668	0.781	0.724	0.493	0.502	$0.633^{\dagger}$

Table 7: Macro F-1 scores using inter-domain adaptation (first block), in-domain temporal adaption (second block), and of our best performing models using Noisy Student training (Xie et al., 2020). We assert significance† if $p < 0.05$ under a paired-t test with base model (CTBERT-GOEMOTIONS for inter-domain transfers and CTBERT-$\mathcal{F}_{tr}$ for temporal transfers.)

in domain temporal performance gaps. First, we note that our SSL-powered CT-BERT model trained on GoEmotions outperforms the plain CT-BERT by as much as 1% in average macro F-1. Moreover, in our temporal transfer experiments, Noisy Student improves the performance of the model by 1.7%. These results are statistically significant, and emphasize that our method obtains better generalization performance and can be leveraged to produces better domain adaptation models.

6 Related Work

Emotion datasets. Emotion detection has been studied extensively with applications in music (Strapparava et al., 2012), social networks (Mohammad, 2012; Islam et al., 2019), online news (Bao et al., 2009), and literature (Liu et al., 2019). All these domains can be examined with the help of large curated datasets. These datasets are created using automated approaches such as distant supervision (Wang et al., 2012; Abdul-Mageed and Ungar, 2017), while others are manually labeled using crowdsourcing (Aman and Szpakowicz, 2007; Poria et al., 2019; Liu et al., 2019; Sosea and Caragea, 2020; Demszky et al., 2020; Desai et al., 2020). In this work, we resort to the latter and create CovidEmo, a dataset of $2,847$ tweets annotated with the Pluchik-8 emotions.

Emotion detection methods. Emotion detection has been studied extensively in the past (Cambria et al., 2017; Stappen et al., 2021; Cambria et al., 2020). In the early stages, most approaches used feature-based methods, which usually leveraged hand-crafted lexicons, such as EmoLex (Mohammad and Turney, 2013) or the Valance Arousal Lexicon (Mohammad, 2018). These features were subsequently used to build classifiers such as Logistic Regression or SVMs. However, due to the recent advancements in deep learning as well as large pre-trained language models, all state-of-the-art approaches (Desai et al., 2020; Sosea and Caragea, 2020; Demszky et al., 2020) employ BERT-based (Devlin et al., 2019) classifiers.

COVID-19 emotion analysis. Since the emergence of the pandemic, numerous studies have been carried out on social media networks to understand COVID-19 and its effects on the larger population. Ils et al. (2021) annotated $2.3K$ German and English tweets for the expression of solidarity and used it to carry out an analysis into the expression of solidarity over time. On the other hand, Saakyan et al. (2021) annotated a dataset for detecting general misinformation in the pandemic. Sentiment analysis and emotion detection on social media during COVID-19 have seen tremendous popularity as well (Beck et al., 2021; Kabir and Madria, 2021; Adikari et al., 2021; Choudrie et al., 2021; Scarpina, 2020; Calbi et al., 2021) due to the ability to provide vital information into the social aspects and the overall dynamics of the population. In this paper, however, we annotate CovidEmo, a dataset of fine-grained emotions and employ a comprehensive analysis into cross-domain and temporal generalization of large pretrained language models. We will make the dataset available to the large public.

7 Conclusion

We present CovidEmo, a dataset of tweets annotated with perceived Plutchik-8 emotions. Using this dataset, we reveal emotion distributions and associations that are distinctive from prior studies on disaster-related emotion annotation and detection. We further show that models trained on other emotion datasets transfer poorly. Additionally, we indicate that models transfer poorly when trained on different temporal slices of an event such as COVID-19. Next, we conduct a comprehensive analysis of the temporal and inter-domain gaps to offer a better understanding of why models transfer poorly. As a potential solution to bridge these gaps and offer a more reliable disaster response, we leverage the large amount of readily available data alongside semi-supervised learning techniques.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728.

Achini Adikari, Rashmika Nawaratne, Daswin De Silva, Sajani Ranasinghe, Oshadi Alahakoon, Damminda Alahakoon, et al. 2021. Emotions of covid-19: Content analysis of self-reported information using artificial intelligence. *Journal of Medical Internet Research*, 23(4):e27341.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205.

Ashwini Ashokkumar and James W Pennebaker. 2021. Social media conversations reveal large psychological shifts caused by covid-19's onset across us cities. *Science advances*, 7(39):eabg7843.

S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*, pages 699–704.

Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in German covid-19 social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–13.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938.

Marta Calbi, Nunzio Langiulli, Francesca Ferroni, Martina Montalti, Anna Kolesnikov, Vittorio Gallese, and Maria Alessandra Umiltà. 2021. The consequences of covid-19 on social interactions: an online study on face covering. *Scientific Reports*, 11(1):1–10.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 105–114, New York, NY, USA. Association for Computing Machinery.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273.

Jyoti Choudrie, Shruti Patil, Ketan Kotecha, Nikhil Matta, and Ilias Pappas. 2021. Applying and understanding an advanced, novel deep learning approach: A covid 19, text based, emotions analysis study. *Information Systems Frontiers*, pages 1–35.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Michael Robert Dennis, Adrianne Kunkel, Gillian Woods, and Paul Schrodt. 2006. Making Sense of New Orleans Flood Trauma Recovery: Ethics, Research Design, and Policy Considerations for Future Disasters. *Analyses of Social Issues and Public Policy*, 6(1):191–213.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Julia Daisy Fraustino, Brooke Fisher Liu, and Yan Xian Jin. 2012. Social Media Use During Disasters: A Review of the Knowledge Base and Gaps. *National Consortium for the Study of Terrorism and Responses to Terrorism*.

Alexandra Ils, Dan Liu, Daniela Grunow, and Steffen Eger. 2021. Changes in European solidarity before and during COVID-19: Evidence from a large crowd- and expert-annotated Twitter dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1623–1637.

Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *IEEE Access*, 8:181074–181090.

Jumayel Islam, Robert E Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1355–1365.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Md. Yasin Kabir and Sanjay Madria. 2021. Emocov: Machine learning for emotion detection, analysis and visualization using covid-19 tweets. *Online Social Networks and Media*, 23:100135.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6293–6298.

Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Saif Mohammad and Svetlana Kiritchenko. 2015. Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Hui Xian Lynnette Ng, Roy Ka-Wei Lee, and Md Rabiul Awal. 2020. I miss you babe: Analyzing emotion dynamics during COVID-19 pandemic. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 41–49, Online. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Benjamin Nye and Ani Nenkova. 2015. Identification and characterization of newsworthy verbs in world news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1440–1445.

Leysia Palen and Kenneth M Anderson. 2016. Crisis informatics-new data for extraordinary times. *Science*, 353(6296):224–225.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129.

Federica Scarpina. 2020. Detection and recognition of fearful facial expressions during the coronavirus disease (covid-19) pandemic in an italian sample: An online experiment. *Frontiers in Psychology*, 11:2252.

Axel Schulz, Tung Dang Thanh, Heiko Paulheim, and Immanuel Schweizer. 2013. A Fine-Grained Sentiment Analysis Approach for Detecting Crisis Related Microposts. In *Information Systems for Crisis Response and Management (ISCRAM)*.

Tiberiu Sosea and Cornelia Caragea. 2020. Canceremo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

Lukas Stappen, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. Muse 2021 challenge: Multimodal emotion, sentiment, physiological-emotion, and stress detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5706–5707.

Carlo Strapparava, Rada Mihalcea, and Alberto Battocchi. 2012. A parallel corpus of music and lyrics annotated with emotions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2343–2346.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring Perceived Demographics from User Emotional Tone and User-Environment Emotional Contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Hyperparameters Used

In all our experiments, we found that a batch size of 16 works best. Additionally, we indicate in Table 8 the best learning rates for our models. We refrain from showing the best learning rates on HurricaneEmo due to low performance, high variance of the results.

B Hyperparameter Search Space

For each emotion, we investigate with batch sizes in the set [8, 16, 32], and train for up to 5 epochs with early stopping. In terms of learning rates, we follow the best practices from the orignal BERT paper and explore learning rates around $5e-5$. Specifically, we experiment with values in the range $1e-5$ -> $9e-5$ with steps of $2e-5$. Although hyperparameter tuning is quite expensive computationally (15 runs per emotion per model), we found that the default BERT setup (5e-5 learning rate and a batch size of 32) works within 0.5% F-1 of the best model.

C Semi-supervised Learning

Noisy Student training (Xie et al., 2020) leverages knowledge distillation (KD) and self-training to iteratively train two models in a teacher-student framework. The framework trains the student in traditional KD fashion, matching its predictions to those of the teacher. Concretely the training loss is:

$$\mathcal{L} = \sum_{(x^i, y^i) \in \mathcal{C}_{tr}} l(f_{\mathcal{T}}(x^i), f_{\mathcal{T}'}(x^i)),$$

where $\mathcal{D}$ is the training dataset, l is the cross-entropy loss, and $f_{\mathcal{T}}$ and $f_{\mathcal{T}'}$ are the student and the teacher models, respectively. We note one vital particularity of this framework: The student is trained using noised input examples. In the orginal paper, the authors also use a larger network for the student, but we noticed here that using equal-sized architectures works well enough. Leveraging noised inputs, Noisy Student exposes the student to more difficult learning environments, and usually leads to an increased performance compared to the teacher. To add noise to our input examples, we use two approaches: a) *Synonym replacement:* We replace between one and three words in a tweet with its synonym using the WordNet English lexical database (Fellbaum, 2012); b) *Back-translation:* We use back-translation, and experiment with different levels of noise corresponding to different translation chain lengths (e.g., English-French-Spanish-English). Smaller chain lengths lead to less noise, while increasing the length of the chain produces examples with significantly more noise. For each unlabeled example, we sample uniformly a chain length in the range 1->10, and use the following languages for translation: Russian, French, Spanish, Italian, and German.

	ang	ant	dis	fea	joy	sad	sur	tru
BERT-GOEMOTIONS	$3e-05$	$5e-05$	$7e-05$	$5e-05$	$5e-05$	$5e-05$	$3e-05$	$5e-05$
BERTWEET-GOEMOTIONS	$3e-05$	$3e-05$	$5e-05$	$5e-05$	$5e-05$	$7e-05$	$7e-05$	$5e-05$
CTBERT-GOEMOTIONS	$5e-05$	$7e-05$	$3e-05$	$5e-05$	$5e-05$	$3e-05$	$3e-05$	$5e-05$
CTBERT-GOEMOTIONS-SSL	$3e-05$	$1e-05$	$5e-05$	$3e-05$	$7e-05$	$7e-05$	$5e-05$	$5e-05$
CTBERT-$\mathcal{F}_{tr}$	$5e-05$	$7e-05$	$5e-05$	$5e-05$	$3e-05$	$3e-05$	$5e-05$	$5e-05$
CTBERT-$\mathcal{F}_{tr}$-SSL	$5e-05$	$5e-05$	$5e-05$	$7e-05$	$7e-05$	$5e-05$	$5e-05$	$5e-05$

Table 8: Best Learning Rates for for our models.