Empathy and Hope: Resource Transfer to Model Inter-country Social Media Dynamics

Clay H. Yoo[♣] Shriphani Palakodety[♡] Rupak Sarkar[♢] Ashiqur R. KhudaBukhsh[♣*]
[♣]Carnegie Mellon University
[♡]Onai
[♢]Maulana Abul Kalam Azad University of Technology
hyungony@andrew.cmu.edu, spalakod@onai.com,
rupaksarkar.cs@gmail.com, akhudabu@cs.cmu.edu

Abstract

The ongoing COVID-19 pandemic resulted in significant ramifications for international relations ranging from travel restrictions, global ceasefires, and international vaccine production and sharing agreements. Amidst a wave of infections in India that resulted in a systemic breakdown of healthcare infrastructure, a social welfare organization based in Pakistan offered to procure medical-grade oxygen to assist India - a nation which was involved in four wars with Pakistan in the past few decades. In this paper, we focus on Pakistani Twitter users' response to the ongoing healthcare crisis in India. While #IndiaNeedsOxygen and #PakistanStandsWithIndia featured among the top-trending hashtags in Pakistan, divisive hashtags such as #EndiaSaySorryToKashmir simultaneously started trending. Against the backdrop of a contentious history including four wars, divisive content of this nature, especially when a country is facing an unprecedented healthcare crisis, fuels further deterioration of relations. In this paper, we define a new task of detecting *supportive* content and demonstrate that existing *NLP for social impact* tools can be effectively harnessed for such tasks within a quick turnaround time. We also release the first publicly available data set[1] at the intersection of geopolitical relations and a raging pandemic in the context of India and Pakistan.

1 Introduction

The COVID-19 pandemic started in late 2019 (Carvalho et al., 2021) and as of this writing is still ongoing. Several factors - geopolitical, economic, social among others - dramatically influenced health outcomes around the world. In this paper, we focus on the ongoing (as of May 2021) infection wave in India (CNN, 2021). After aggressive initial steps to successfully curb the spread of the virus, case counts exploded in India towards the end of April 2021. The rapidity of the spread overwhelmed the healthcare infrastructure in the country. A widespread shortage of medical-grade oxygen (BBC, 2021), overworked medical staff, and full capacity emergency rooms became the norm in major population centers.

The crisis was heavily discussed on social media and the associated hashtags were among the most discussed Twitter trends globally. In Pakistan, a neighboring country that fought four wars with India over the past seven decades (Paul and Paul, 2005), a significant volume of tweets expressed solidarity with the Indian populace primarily through two hashtags - #IndiaNeedsOxygen and #PakistanStandsWithIndia. In addition, the hashtag #EndiaSaySorryToKashmir started trending in Pakistan. The tweets using this hashtag were primarily divisive and often referenced a long-running territorial dispute at the heart of India-Pakistan relations. Amidst a far-reaching and rapidly progressing pandemic, divisive content of this nature negatively impacts the mental well-being of the affected population and can contribute to strained relations.

Hashtag based filtering, while extremely effective, cannot solely identify *supportive* content. For instance, users can hijack trending hashtags and post content that violates the spirit of the hashtag (see Table 1). Also, replies or responses to a controversial tweet with a divisive hashtag may still retain the same hashtag but the content may reflect a unifying message. Rapidly evolving crises also require a fast turnaround time which can preclude

[*] Ashiqur R. KhudaBukhsh is the corresponding author.
[1] Data is publicly available at https://github.com/anton-sturluson/empathy-and-hope.

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 125–134
Bangkok, Thailand (online), August 5, 2021. ©2021 Association for Computational Linguistics

#PakistanStandsWithIndia	we're rivals not enemies. we breath same air speak same languages. our prayers , wishes and thoughts are with our brothers from other side of the border. We need to fight this bettle together
#PakistanstandswithIndia	karma is bitch, india deserves what's happening right now because that's what they did with people of kashmir. kashmir's can't take revenge but god has his plans for redemption.
#IndiaNeedsOxygen	Despite the fact that we have our political conflicts, but I really pray for their good health. Get well soon india. Pakistani nation is with you.
#IndiaNeedsOxygen	India deserves this . You are facing what you did to kashmir and fool pakistani supporting india on this you are just slaves to british thats all ..
#EndiaSaySorryToKashmir	Kashmir is our and it is all of it. Until the independence of Kashmir, there will be war till the destruction of India.
#EndiaSaySorryToKashmir	Political differences have their place but the prayers of us Pakistanis are with our Indian brothers and sisters. May Allah give health to all.

Table 1: Example tweets where the hashtag and the tweet content agree (highlighted in blue) and disagree (highlighted in red).

sophisticated, time-consuming solutions.

In this paper, we present a method to automatically detect *supportive* content from the tweet text (excluding hashtags, mentions, emojis, and urls). Our minimally supervised approach combines multiple soft signals - a *hope speech* classifier that detects peace-seeking content (Palakodety et al., 2020a), and an *empathy-distress* classifier trained on a well-known empathy-distress data set (Buechel et al., 2018). We further demonstrate superior performance in presence of supervision and release an annotated data set in this important humanitarian domain.

Model reusability is a major challenge in NLP applications (Arango et al., 2019; Beltagy et al., 2019). We see our paper as preliminary evidence that NLP methods for positive impact research are not isolated efforts, and solutions arising from adjacent tasks can be re-purposed to tackle newer challenges.

NLP for positive impact: Our work can be described by the following two broad themes specific to this workshop - *online well-being & positive information sharing* and *case studies for NLP for social good*. In order to create a positive impact, we believe a research contribution needs to satisfy a subset of the following conditions: (1) a problem domain with a high societal impact; (2) resource-sharing to facilitate scientific progress; and (3) a research theme that spawns a rich line of follow-up work.

Our paper has the following contributions:

Social: We analyze the bilateral relationship between countries with a contentious history amidst a raging pandemic. Our work is at the intersection of two important themes - geopolitical relations and healthcare crises. We show a significant outpouring of support and solidarity between the two nations' online communities in the context of the pandemic. Barring a few recent efforts (Palakodety et al., 2020a; Tyagi et al., 2020), there is little literature on web manifestation of the India-Pakistan relationship co-occurring with other crises. To the best of our knowledge, this is the first analysis of social media text interactions between India and Pakistan amidst a pandemic.

Resource: We present a data set of tweets exploring geopolitical relations between historic adversaries amidst a health crisis. Publicly available data sets expressing empathy and distress are scarce (Buechel et al., 2018). Beyond our immediate objective of detecting *supportive* tweets, this data set may be useful in answering several other research questions.

The reusability argument: We present a compelling case study that *NLP for positive impact applications* are not isolated tasks. Rather, multiple existing resources can be combined to tackle a new challenge in a fast turnaround time setting.

2 Task

In this paper, we consider the task of detecting *supportive* content. Supportive behavior in language has been previously studied. For example, a AAAI-2020 shared task focused on detecting *disclosure* and *supportiveness* from written accounts of casual and confessional conversations (Chhaya et al., 2020). Our task is slightly different in the sense that we are interested in detecting content where speakers are supporting a country/people severely affected by a healthcare crisis.

We define *supportive* content to be either expressing empathy, distress, or solidarity. Our def-

Empathy	Our hearts go out to our neighbours who are facing unprecedented misery. Pakistani People are praying for you . . .
Distress	I am a Pakistani but seriously this is heartbreaking what i am seeing from few days about India.We are enemies but this is about humanity,If we unite in this pandemic we both countries can fight together and can win this battle together,Peace . . .
Solidarity	As a human we all are together Pray for India and for all people all over the world who are suffering from COVID May Allah pak save us from this dangerous COVID-19 Stop hating start praying

Table 2: Example tweets exhibiting empathy, distress, and solidarity.

initions for empathy and distress follow (Buechel et al., 2018) that considers extensive psychology literature (Batson et al., 1987; Batson and Shaw, 1991; Sober and Wilson, 1999; Goetz et al., 2010; Mikulincer and Shaver, 2010). (Buechel et al., 2018) defines empathy as a warm, tender, and compassionate feeling for a suffering entity, and distress as a self-focused, negative affective state that occurs when one feels upset due to witnessing an entity's suffering or need. Among the several existing definitions of solidarity, we borrow the following (Wildt, 1999): a mutual attachment between individuals (groups) that encompasses two levels: (1) a *factual level* of actual common ground between the individuals (groups); and (2) a *normative level* of mutual obligations to aid each other, as and when should be necessary. In Table 2 we present three example tweets exhibiting empathy, distress, and solidarity.

Our definition for *not-supportive* content does not have a similar psychological grounding. Our annotators observed that the *not-supportive* content in this specific context, primarily (1) expressed politically motivated hate; (2) demonstrated a warmongering attitude; (3) expressed schadenfreude; (4) mentioned politically contentious issues; and (5) expressed unrelated content such as product promotion etc.

3 Resource

We use two existing resources for our work. Next, we present a short description of these resources.

3.1 *Hope speech* classifier

The *hope speech* detection task introduced in (Palakodety et al., 2020a) involves identifying social media text content with a unifying message encouraging peace, discouraging war, and highlighting the economic, social, and human costs of conflict against the backdrop of the 2019 India-Pakistan conflict. A detailed definition of *hope speech* with illustrative examples is provided in (Palakodety et al., 2020a).

3.2 Empathy and Distress Classifier

We train a classifier on the empathy-distress data set introduced in (Buechel et al., 2018). The data set is grounded in prior psychology literature on empathy and distress (Batson et al., 1987; Batson and Shaw, 1991; Sober and Wilson, 1999; Goetz et al., 2010; Mikulincer and Shaver, 2010). The data set consists of 418 news article excerpts from popular news platforms and responses to them from 403 annotators, resulting in a total of 2,015 responses (5 articles per annotator). Upon filtering responses that deviated from the task description, the pruned final data set consists of 1,860 responses (empathy: 916, distress: 905). We split this data into train and test sets in 90/10 ratio and train a binary classifier using `BERT` (Devlin et al., 2019) (`bert-base-uncased`) using transformers library (Wolf et al., 2020).

4 Data

Our data set, $\mathcal{T}$, consists of 309,394 tweets posted by 150,289 unique users collected between 21 April 2021 and 04 May 2021. The top trending hashtags in Pakistan for April 22 and April 23 were retrieved from `https://getdaytrends.com/` and all associated tweets were obtained using the Twitter API[2]. Other closely related trending hashtags were also included (e.g., #IndiaNeedsOxygen and #IndiaNeedOxygen, or #PakistanStandsWithIndia and #PakistanStandWithIndia). Additional details are in Table 4. In this paper, any mention of a hashtag includes closely spelled variants (e.g. #IndiaNeed(s)Oxygen, #PakistanStand(s)WithIndia, or #I(E)ndiaSaySorryToKashmir). We define the following two hashtag sets: $\mathcal{H}_{supportive}$ = {#IndiaNeed(s)Oxygen, #PakistanStand(s)WithIndia} ; and $\mathcal{H}_{not\text{-}supportive}$ = {#I(E)ndiaSaySorryToKashmir}.

Subsets of interest: Two mutually disjoint subsets of $\mathcal{T}$: $\mathcal{T}_{supportive}$ and $\mathcal{T}_{not\text{-}supportive}$ are de-

[2] https://developer.twitter.com/en/docs/twitter-api

fined as follows. $\mathcal{T}_{supportive}$ includes tweets containing one or more of the $\mathcal{H}_{supportive}$ hashtags and $\mathcal{T}_{not\text{-}supportive}$ includes tweets containing one or more of the $\mathcal{H}_{not\text{-}supportive}$ hashtags. Tweets containing any intersection of the $\mathcal{H}_{supportive}$ and $\mathcal{H}_{not\text{-}supportive}$ hashtags are discarded from either subset and thus there is no intersection between $\mathcal{T}_{supportive}$ and $\mathcal{T}_{not\text{-}supportive}$. Since classification of extremely short texts is a well-established challenge (Sindhwani et al., 2009; Attenberg et al., 2010; KhudaBukhsh et al., 2015), in all of our sampling experiments involving a text classifier, we impose a length restriction of 10 or more tokens after preprocessing. Furthermore, our classifiers are only presented with the tweet text, i.e., the body of the tweet with hashtags, emojis, urls, and mentions removed.

Generating country labels for tweets: The Twitter API bundles geographic location (coordinates) with tweets. In addition, we utilized a weak signal - if a user's Twitter handle contains an India or Pakistan flag emoji, then we assume their tweets originated in India or Pakistan respectively. In the cases where the location information and our signal are both present, we notice no inconsistency, indicating our weak country signal is robust.

5 Characterization of the Tweets

5.1 Likes and Retweets

We now characterize the retweets and likes of each of these hashtags. Let $\#ht_{Ind}$, $\#ht_{Pak}$, and $\#ht_{Other}$ denote the subsets of tweets that contain the hashtag ht and originate in India, Pakistan, and other (or unknown), respectively. Table 5 shows that overall, the tweets containing *supportive* hashtags received fewer likes and retweets than those containing *not-supportive* hashtags. We further notice that tweets containing *supportive* hashtags that originated in Pakistan received substantially more likes than those from India. Our results though come with the following caveats. Multiple factors can influence our data collection process such as the inner workings of Twitter algorithms or the Twitter API. Also, our focus is on English tweets; previous studies have reported that Hindi is more commonly used to express negative sentiment in social media content generated in the Indian subcontinent (Rudra et al., 2016; KhudaBukhsh et al., 2020).

5.2 Hashtag Co-occurrence

We next measure in-group and out-group co-occurrence of *supportive* and *not-supportive* hashtags within a single tweet. Pair-wise Jaccard index between the tweet sets using various hashtags is computed[3] and shown in Table 3. We observe that among all hashtag pairs, ⟨#IndiaNeed(s)Oxygen and #PakistanStand(s)WithIndia⟩ occurs the most. We observe that qualitatively, there is a stark contrast between tweets containing $\mathcal{H}_{supportive}$ hashtags and tweets containing $\mathcal{H}_{not\text{-}supportive}$ hashtags with the dominant theme in the former being empathy, distress, and solidarity. Figure 1 presents a word-cloud visualization of the tweets employing the three hashtags.

6 Related Work

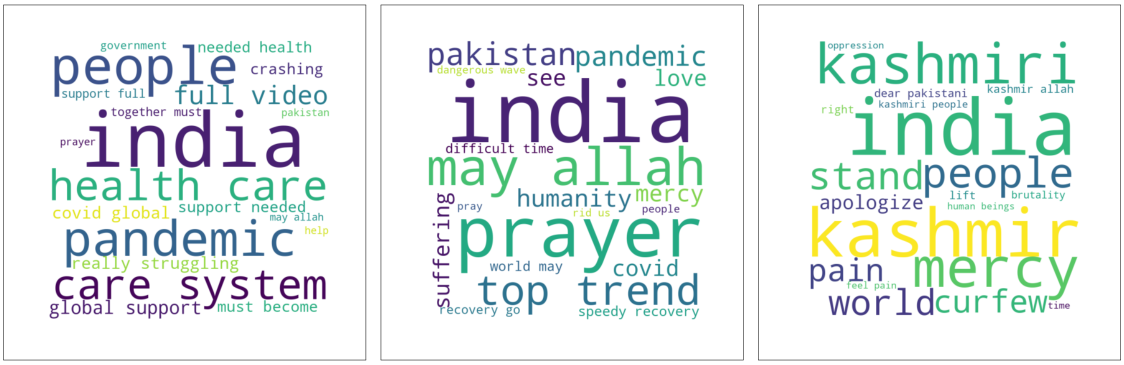
Social media response to the ongoing pandemic has received significant research attention: (1) health misinformation (Memon and Carley, 2020; Hossain et al., 2020; Cinelli et al., 2020), (2) polarization (Cruickshank and Carley, 2020; KhudaBukhsh et al., 2021), (3) disease modeling (Li et al., 2020), etc. Counterhate measures along the line of counterspeech research (Benesch et al., 2016; Benesch, 2014; Mathew et al., 2018; Palakodety et al., 2020b) to combat Anti-Asian hate (Ziems et al., 2020), and community blame (Saha et al., 2021) has been studied. Our work contrasts with existing literature in three ways: (1) we analyze bilateral relations of nuclear adversaries amidst a raging pandemic; (2) we release a novel data set for wider use exploring related research questions; and (3) we present a new method that combines recent *NLP for positive impact* advances in a new, timely, and important task.

While the political volatility between India and Pakistan has been extensively studied by social scientists (Malik and Wirsing, 2002; Schofield, 2010; Bose, 2009), barring few recent lines of work (Palakodety et al., 2020a; KhudaBukhsh et al., 2020; Tyagi et al., 2020), social media interactions between the civilians of India and Pakistan has received little or no attention. All recent work on Indian and Pakistani social media (Palakodety et al., 2020a; KhudaBukhsh et al., 2020; Tyagi et al., 2020) focused on a solitary incident - the 2019 India-Pakistan conflict triggered by the Pulwama terror attack across different social media platforms.

[3] Jaccard index is a statistic to gauge similarity between two sets, $\mathcal{A}, \mathcal{B}$, expressed as $\frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}$.

hashtags	#IndiaNeed(s)Oxygen	#PakistanStand(s)WithIndia	#I(E)ndiaSaySorryToKashmir
#IndiaNeed(s)Oxygen	-	0.0887	0.0247
#PakistanStand(s)WithIndia	0.0887	-	0.0405
#I(E)ndiaSaySorryToKashmir	0.0247	0.0405	-

Table 3: Jaccard index of tweet subsets employing various hashtags.



(a) #IndiaNeed(s)Oxygen (b) #PakistanStand(s)WithIndia (c) #I(E)ndiaSaySorryToKashmir

Figure 1: A word cloud visualization of the tweet contents and the associated hashtag used. Hashtags and punctuations are removed as a preprocessing step.

Hashtag	Total	India	Pakistan
#IndiaNeedsOxygen	145,975	26,383	19,748
#IndiaNeedOxygen	24,488	5,049	2,400
#PakistanStandsWithIndia	96,226	12,331	21,583
#PakistanStandWithIndia	17,406	2,772	3,790
#EndiaSaySorryToKashmir	25,081	87	8,022
#IndiaSaySorryToKashmir	557	15	169
All	309,733	46,651	55,712

Table 4: Statistics of dataset crawled between 21 April 2021 and 04 May 2021.

While (Palakodety et al., 2020a) introduced a novel task of detecting hostility-diffusing, peace seeking *hope speech* and considered comments on relevant YouTube videos as the data set, (Tyagi et al., 2020) is the first work on analyzing web-manifestation (Twitter) of political polarization between the two countries and how political parties factor in these discussions.

Our work leverages two existing resources: (1) a *hope speech* classifier introduced in (Palakodety et al., 2020a); and (2) a well-known *empathy-distress* data set (Buechel et al., 2018). As already mentioned, our work differs in a key way that we re-purpose these resources for a new *NLP for positive impact* task: detecting *supportive* tweets in the context of social media discussions during a national healthcare crisis. Our work also draws inspiration from recent findings about mining stance from hashtags (Kumar, 2018).

7 Methods, Results, and Discussion

Research question: *Does sampling tweets containing $\mathcal{H}_{supportive}$ hashtags alone suffice?*

We first investigate if hashtag-based filtering alone guarantees *supportive* tweets with a high probability. We randomly sample 1,000 tweet texts from $\mathcal{T}_{supportive}$ and manually annotate them. Our annotators are provided only the tweet texts, i.e., the body of the tweet excluding hashtags, urls, mentions, and emojis. Three annotators fluent in English, Hindi, and Urdu, and well-versed with the geopolitical events between India and Pakistan first independently annotated these tweets and achieved a Fleiss' κ score of 0.76 indicating moderate agreement. Next, disagreements are resolved through a follow-up adjudication process and a higher Fleiss' κ score of 0.86 is reached. Of the randomly chosen 1,000 tweets 444 tweets, i.e., 44.4% were marked positive. This result indicates that solely relying on *supportive* hashtag will not do better than chance and underscores the importance of sophisticated methods.

In addition, we randomly sampled 1,000 tweet texts from $\mathcal{T}_{supportive} \cup \mathcal{T}_{not\text{-}supportive}$ as our test set (denoted as $\mathcal{D}_{eval}$). Throughout our annotation process, whenever consensus label is absent, following standard literature (Bowman et al., 2015), we consider the majority label as the gold-standard label. Annotator subjectivity is a well-studied research

Hashtag$_{Location}$	Like	Retweet
#IndiaNeed(s)Oxygen$_{Ind}$	2.32 ± 63.80	1631.07 ± 3393.86
#IndiaNeed(s)Oxygen$_{Pak}$	4.39 ± 96.50	322.98 ± 1107.28
#IndiaNeed(s)Oxygen$_{Other}$	2.72 ± 215.81	1306.71 ± 2934.60
#PakistanStand(s)WithIndia$_{Ind}$	2.46 ± 78.14	2313.45 ± 2898.67
#PakistanStand(s)WithIndia$_{Pak}$	8.58 ± 358.16	665.03 ± 1559.59
#PakistanStand(s)WithIndia$_{Other}$	2.65 ± 117.25	1246.58 ± 2195.85
#I(E)ndiaSaySorryToKashmir$_{Ind}$	1.49 ± 4.97	191.45 ± 266.38
#I(E)ndiaSaySorryToKashmir$_{Pak}$	1.26 ± 24.80	276.28 ± 300.87
#I(E)ndiaSaySorryToKashmir$_{Other}$	1.51 ± 37.61	248.33 ± 293.02

Table 5: Location-specific like and retweet behavior.

Label	Percentage	Like	Retweet
supportive$_{Pak}$	85.30%	6.64 ±270.6	505.61 ±1378.1
not-supportive$_{Pak}$	14.70%	1.26 ± 24.8	276.28 ± 300.9

Table 6: Like and retweet behavior and count of *supportive* and *not-supportive* tweets from Pakistan.

Model	Precision	Recall	F1
$\mathcal{M}^{\text{BERT}}_{supervised}$	83.28 ± 0.8	80.98 ± 1.6	81.14 ± 1.6
$\mathcal{M}^{\text{BERT}}_{informed}$	80.78 ± 0.5	80.60 ± 0.7	80.62 ± 0.6
$\mathcal{M}^{\text{BERT}}_{hashtag}$	72.93 ± 1.2	53.78 ± 1.6	48.58 ± 2.5
$\mathcal{M}^{\text{SVM}}_{supervised}$	66.38 ± 0.5	91.65 ± 0.8	76.99 ± 0.4
$\mathcal{M}^{\text{SVM}}_{informed}$	56.98 ± 0.6	94.03 ± 0.3	70.95 ± 0.5
$\mathcal{M}^{\text{SVM}}_{hashtag}$	42.69 ± 0.03	100.00 ± 0	59.83 ± 0.03

Table 7: Test performance comparison. Five runs per experiment were conducted and mean and standard deviation are presented.

area (Pavlick and Kwiatkowski, 2019), and in order to facilitate further research, we also provide individual annotator's labels.

Research question: *Do the hope speech and the empathy-distress classifiers present any discernible signal to differentiate between supportive and not-supportive tweets?*

As already described, the *hope speech* classifier is designed for a different scenario of detecting peace-seeking, hostility diffusing content from social media discussions generated during a conflict. Our current task of detecting *supportive* tweets, although related, is not identical. Furthermore, the classifier is trained on a different social media platform, YouTube, that allows unstructured text without any length restriction, whereas Twitter allows unstructured text but imposes a length restriction. Similarly, the *empathy-distress* classifier is trained on a different data set of user responses to news events. Hence, a pertinent research question is if the *hope speech* classifier or the *empathy-distress* classifier is any good in differentiating between *supportive* and *not-supportive* tweets.

We first start with a simple experiment to illustrate that the resources provide useful signal. Let $S = \{\langle x, y\rangle\}$ such that $x \sim \mathcal{T}_{supportive}$ and $y \sim \mathcal{T}_{not\text{-}supportive}$, i.e., S consists of tweet pairs $\langle x, y\rangle$ where x and y are randomly drawn from the pool of tweets with *supportive* and *not-supportive* hashtags, respectively. Let $\mathcal{P}_h(z)$ and $\mathcal{P}_e(z)$ denote the predicted *hope speech* and *empathy-distress* probabilities of tweet z. We compute: $r_h = \frac{\Sigma_{\langle x,y\rangle \in \mathcal{S}_s} \mathbb{I}(\mathcal{P}_h(x) > \mathcal{P}_h(y))}{|\mathcal{S}|}$ and $r_e = \frac{\Sigma_{\langle x,y\rangle \in \mathcal{S}_s} \mathbb{I}(\mathcal{P}_e(x) > \mathcal{P}_e(y))}{|\mathcal{S}|}$ where $\mathbb{I}$ denotes an indicator function and $|\mathcal{S}|$, i.e., the number of randomly drawn pairs, is set to 100,000. We ran this experiment five times and found r_h to be equal to 69.3 ± 0.13% and r_e to be equal to 47.8 ± 0.12%, indicating that a randomly drawn sample from $\mathcal{T}_{supportive}$ is more likely to receive a higher *hope speech* score ($\mathcal{P}_h(.)$) than a randomly drawn sample from $\mathcal{T}_{not\text{-}supportive}$. However, we do not notice similar trends with our *empathy-distress* classifier.

It is unsurprising that r_h has a much higher value than r_e. The *hope speech* classifier is trained on a data set relevant to a recent India-Pakistan conflict and thus has a substantial overlap in domain. Hence, a general nature of positive dialogue may indicate a desire to put things behind and help each other. In contrast, the *empathy-distress* classifier is trained on a broad, diverse, data set of user responses to news events and has no overlap with the current domain. However, when we rank tweets from $\mathcal{T}_{supportive}$ by the classifier's probability, we notice that top predictions are of extremely high quality in both cases. We annotate top 1,000 unique tweets from $\mathcal{T}_{supportive}$ ranked by $\mathcal{P}_h(.)$ and obtain 950 positives. Similarly, top 1,000 unique tweets from $\mathcal{T}_{supportive}$ ranked by $\mathcal{P}_e(.)$ yield 899 positives upon manual annotation. Moreover, the two classifiers complement each other as among the top 1,000 unique tweets from the *hope speech* classifier and the top 1,000 unique tweets from the *empathy-*

distress classifier had minimal overlap (62 samples). This annotation task also yielded a substantially higher Fleiss' κ score (0.8068) without any follow-up adjudication process indicating that the chosen samples have less ambiguity than our earlier experiment that involved annotating randomly selected tweets from $\mathcal{T}_{supportive}$. Our results thus indicate existing resources can be harnessed for informed sampling yielding high-quality positives.

Research question: *How to leverage existing resources to design an effective classifier to detect supportive tweets?*

We utilize two existing resources, a *hope speech* classifier from (Palakodety et al., 2020a), and an *empathy-distress* data set from (Buechel et al., 2018). We first train an *empathy-distress* classifier on the *empathy-distress* data set that can classify tweets as exhibiting empathy or distress, or not.

Our pipeline utilizes the *hope speech* and *empathy-distress* classifiers and constructs a weakly labeled data set where the positive examples exhibit themes like empathy, distress, support, and solidarity - the *supportive speech*, and the negative examples exhibit themes like controversy, whataboutism, and hostility - the *not-supportive speech*. The two classifiers are used to label tweets and the positive class probability is used to rank all the tweets in the set $\mathcal{T}_{supportive} \cup \mathcal{T}_{not\text{-}supportive}$ yielding two ranked lists. $\mathcal{D}^{+}{}_{informed}$ contains all tweets using any of the top 1,000 tweets in both ranked lists (2,000 in total, 1,938 unique) are considered positive samples, and a set of negative samples, $\mathcal{D}^{-}{}_{informed}$, is constructed by randomly sampling 500 tweets each from the bottom 80% of both ranked lists (1,000 in total, 1,000 unique). The full data set construction pipeline is presented in Algorithm 1. The trained model is denoted as $\mathcal{M}_{informed}$.

Earlier research has reported hashtags as an effective way to obtain weak labels (Kumar, 2018). We contrast $\mathcal{M}_{informed}$ against a baseline that uses hashtags alone as a source of weak labels and contains the identical number of (weakly labeled) positives and negatives as $\mathcal{D}_{informed}$. Essentially, any tweet belonging to $\mathcal{T}_{supportive}$ is considered a positive and any tweet belonging to $\mathcal{T}_{not\text{-}supportive}$ is considered a negative. Positives and negative examples are randomly sampled from these sets and a data set with the same proportions as $\mathcal{D}_{informed}$ is constructed. The trained model is denoted as $\mathcal{M}_{hashtag}$.

We train our classifiers using `BERT` (Devlin et al., 2019) (`bert-base-uncased`) using the transformers library (Wolf et al., 2020) and a 90/10 train/validation split. In addition, since English social media content from the Indian subcontinent exhibits a variety of disfluencies (Sarkar et al., 2020), and since the SVM baseline has been successfully applied to the original *hope speech* detection task (Palakodety et al., 2020a), we include an SVM baseline as well that uses TF-IDF vectors as document feature representations. The trained models are evaluated on $\mathcal{D}_{eval}$, 1000 randomly sampled tweets from $\mathcal{T}_{supportive} \cup \mathcal{T}_{not\text{-}supportive}$. Note that hashtags, urls, emojis, mentions, and punctuation are removed from the tweets prior to training.

7.1 Performance Comparison

Table 7 shows that $\mathcal{M}_{informed}$ substantially outperforms $\mathcal{M}_{hashtag}$ on the test set and thus underscores why hashtag-based-filtering may not solely suffice. Also, this result indicates that the joint concept of empathy, distress, and solidarity is learnable, and in this context, the resources exhibit synergy. Understandably, a supervised solution will improve the performance since weak labels obtained using the *hope speech* and *empathy-distress* classifier, while high-quality, still had some amount of noise. Compared to the informed sampling, we observe a slight performance boost in our supervised solutions. We also notice the `BERT`-based classifiers outperformed SVM baselines.

While our primary focus is on Twitter, several social media platforms exist where hashtags are not as prevalent. YouTube, a highly popular social media platform, is one such example. We performed an in-the-wild test where we obtained the top 100 *supportive* predictions from a new data set consisting of 31,232 comments on 185 YouTube COVID-19-related videos from the official YouTube channel of Geo TV, a highly popular Pakistani news channel. We used the best $\mathcal{M}^{\mathrm{BERT}}_{informed}$ model to test our minimally supervised method's in-the-wild performance. Out of 100 such comments, a manual evaluation revealed that 70 were positive. Table 8 lists a few such randomly sampled comments. A reasonably high precision of our model indicates its cross-platform viability and applicability in downstream tasks like moderation.

7.2 Discussion

Research question: *How Pakistan Responded to this crisis?* In our earlier analysis in Section 5.1, we

```
Algorithm 1: $Construct(\mathcal{D}_{informed}, \mathcal{M}_{informed})$
Input: $\mathcal{T}$ is the full set of tweets, $\mathcal{T}_{supportive}, \mathcal{T}_{not\text{-}supportive} \subset \mathcal{T}$; $\mathcal{M}_{hopeSpeech}$ is the hope speech classifier;
  $\mathcal{M}_{empathyDistress}$ is the empathy-distress classifier
Output: $\mathcal{D}_{informed} \subset \mathcal{T}$; and $\mathcal{M}_{informed}$ - a model trained on $\mathcal{D}_{informed}$
Procedure:
foreach tweet $t \in \mathcal{T}_{supportive} \cup \mathcal{T}_{not\text{-}supportive}$ do
    classify $t$ using $\mathcal{M}_{hopeSpeech}$ and $\mathcal{M}_{empathyDistress}$ yielding positive probabilities $\mathcal{P}_h$ and $\mathcal{P}_e$.
end
Sort $\mathcal{T}_{supportive}$ using $\mathcal{P}_h$ and $\mathcal{P}_e$ yielding two ranked lists $\mathcal{R}_{supportive_h}$ and $\mathcal{R}_{supportive_e}$.
Take the top 1,000 tweets from $\mathcal{R}_{supportive_h}$ and $\mathcal{R}_{supportive_e}$ yielding 2,000 tweets - these are the positive samples -
  $\mathcal{D}^{+}{}_{informed}$.
Sort $\mathcal{T}_{not\text{-}supportive}$ using $\mathcal{P}_h$ and $\mathcal{P}_e$ yielding two ranked lists $\mathcal{R}_{not\text{-}supportive_h}$ and $\mathcal{R}_{not\text{-}supportive_e}$.
Sample 500 tweets from the bottom 80% of $\mathcal{R}_h$ and $\mathcal{R}_e$ yielding 1,000 tweets - these are the negative samples -
  $\mathcal{D}^{-}{}_{informed}$.
$\mathcal{D}_{informed} \leftarrow \mathcal{D}^{+}{}_{informed} \cup \mathcal{D}^{-}{}_{informed}$
Duplicates are discarded from $\mathcal{D}_{informed}$
$\mathcal{M}_{informed} \leftarrow$ a classifier trained on $\mathcal{D}_{informed}$
Output: $\mathcal{D}_{informed}$ and $\mathcal{M}_{informed}$
```

Life is dying in our neighboring country. We have differences. We have fought wars, but we are neighbors. Sighing lives in India. My lord, who will do good except you There is no religion of humanity. May Allah save the whole world including India from this epidemic. Amen
From Pakistan I request my all Muslims Humanity has no religion and no boundaries ....Pray for all world and for India
Be safe everone, wear mask everytime, may your country doesn't goes through what our country is going. Greetings from india

Table 8: Randomly sampled YouTube comments predicted as *supportive* by $\mathcal{M}^{\texttt{BERT}}_{informed}$ in the wild.

found that tweets containing $\mathcal{H}_{supportive}$ hashtags originating in Pakistan (1) heavily outnumbered those containing $\mathcal{H}_{not\text{-}supportive}$ hashtags; and (2) received a larger share of the likes and retweets. We investigate the like and retweet behavior conditioned on the tweet text less the hashtags. Table 6 indicates an overwhelming majority of the tweets from Pakistan is classified as *supportive* by $\mathcal{M}_{supervised}$ and such tweets received substantially more likes and retweets than the *not-supportive* tweets.

8 Ethical and Societal Implications

While the setting discussed in the paper involves humanitarian tasks, the techniques can be trivially adapted with the explicit objective to censor empathetic content. In many recent conflicts in the Indian subcontinent, such systems can have adverse social effects, and thus particular care is needed before these systems are deployed. Also, language-specific features can sometimes cause syntactically similar but semantically opposite content to be surfaced underscoring the need for a human-in-the-loop setting before such systems are deployed for social media content moderation tasks. Finally, our classifier relies on a black box *hope speech* classifier and thus runs the risk of propagating possible biases from the black box model. Further case studies need to be considered before deployment and we welcome a thorough investigation of our released data set from the scientific community.

9 Conclusions

In this paper, we present a task and associated resources for a vital domain - geopolitical relations against the backdrop of a raging pandemic. We release a data set of tweets discussing the oxygen crisis and healthcare system collapse in India due to a COVID-19 wave. Our data set is geographically diverse and connects several diverse themes - a long acrimonious history between two neighboring countries that involves four wars and a recent bilateral relations breakdown, a raging pandemic that has claimed several hundred thousand lives within a few weeks and is still ongoing. Our analysis reveals a strong humanitarian streak that prioritizes health and well-being over past geographical or ethnic disputes. We then re-purpose existing resources designed for adjacent tasks like *hope speech* and *empathy distress* detection and utilize these to identify *supportive* tweets. Our experiments reveal that *NLP for positive impact* tasks can utilize existing adjacent resources to rapidly bootstrap solutions.

References

2021. Covid: India sees world's highest daily cases amid oxygen shortage. https://www.bbc.com/news/world-asia-india-56826645. Online; accessed 7-June-2021.

2021. India is spiraling deeper into covid-19 crisis. here's what you need to know. https://www.cnn.com/2021/04/26/india/india-covid-second-wave-explainer-intl-hnk-dst/index.html. Online; accessed 7-June-2021.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Josh Attenberg, Prem Melville, and Foster Provost. 2010. A unified approach to active dual supervision for labeling features and examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 40–55. Springer.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

C Daniel Batson and Laura L Shaw. 1991. Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological inquiry*, 2(2):107–122.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Susan Benesch. 2014. Defining and diminishing hate speech. *State of the World's Minorities and Indigenous Peoples*, 2014:18–25.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on twitter: A field study. *A report for Public Safety Canada under the Kanishka Project*.

Sumantra Bose. 2009. *Kashmir: Roots of conflict, paths to peace*. Harvard University Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Thiago Carvalho, Florian Krammer, and Akiko Iwasaki. 2021. The first 12 months of covid-19: a timeline of immunological insights. *Nature Reviews Immunology*, 21(4):245–256.

Niyati Chhaya, Kokil Jaidka, Lyle Ungar, Jennifer Healey, and Atanu Sinha. 2020. Editorial for the 3rd aaai-20 workshop on affective content analysis.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Scientific Reports*, 10(1):1–10.

Iain J. Cruickshank and Kathleen M. Carley. 2020. Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering. *Appl. Netw. Sci.*, 5(1):66.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Jennifer L Goetz, Dacher Keltner, and Emiliana Simon-Thomas. 2010. Compassion: an evolutionary analysis and empirical review. *Psychological bulletin*, 136(3):351.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Ashiqur R KhudaBukhsh, Paul N Bennett, and Ryen W White. 2015. Building effective query classifiers: a case study in self-harm intent detection. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1735–1738.

Ashiqur R. KhudaBukhsh, Shriphani Palakodety, and Jaime G. Carbonell. 2020. Harnessing code switching to transcend the linguistic barrier. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4366–4374. ijcai.org.

Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. 2021. We don't speak the same language: Interpreting polarization through machine translation. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, page To Appear. AAAI Press.

Sumeet Kumar. 2018. Weakly supervised stance learning using social-media hashtags.

Cuilian Li, Li Jia Chen, Xueyu Chen, Mingzhi Zhang, Chi Pui Pang, and Haoyu Chen. 2020. Retrospective analysis of the possibility of predicting the covid-19 outbreak from internet searches and social media data, china, 2020. *Eurosurveillance*, 25(10):2000199.

Iffat Malik and Robert G Wirsing. 2002. *Kashmir: Ethnic conflict international dispute*. Oxford University Press Oxford.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Shahan Ali Memon and Kathleen M. Carley. 2020. Characterizing COVID-19 misinformation communities using a novel twitter dataset. In *Proceedings of the CIKM 2020 Workshops co-located with 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), Galway, Ireland, October 19-23, 2020*, volume 2699 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mario Ed Mikulincer and Phillip R Shaver. 2010. *Prosocial motives, emotions, and behavior: The better angels of our nature.* American Psychological Association.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020a. Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1881–1889. IOS Press.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020b. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 454–462. AAAI Press.

Thazha Varkey Paul and Thazha Varkey Paul. 2005. *The India-Pakistan conflict: an enduring rivalry*. Cambridge University Press.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas. Association for Computational Linguistics.

Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. ”short is the road that leads from fear to hate”: Fear speech in indian whatsapp groups. *CoRR*, abs/2102.03870.

Rupak Sarkar, Sayantan Mahinder, and Ashiqur KhudaBukhsh. 2020. The non-native speaker aspect: Indian English in social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 61–70, Online. Association for Computational Linguistics.

Victoria Schofield. 2010. *Kashmir in conflict: India, Pakistan and the unending war*. Bloomsbury Publishing.

Vikas Sindhwani, Prem Melville, and Richard D Lawrence. 2009. Uncertainty sampling and transductive experimental design for active dual supervision. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 953–960.

Elliot Sober and David Sloan Wilson. 1999. *Unto others: The evolution and psychology of unselfish behavior*. 218. Harvard University Press.

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. 2020. A computational analysis of polarization on indian and pakistani social media. In *Social Informatics - 12th International Conference, SocInfo 2020, Pisa, Italy, October 6-9, 2020, Proceedings*, volume 12467 of *Lecture Notes in Computer Science*, pages 364–379. Springer.

Andreas Wildt. 1999. Solidarity: its history and contemporary definition. In *Solidarity*, pages 209–220. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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. *CoRR*, abs/2005.12423.