Social Media Attributions in the Context of Water Crisis

Rupak Sarkar♣* **Sayantan Mahinder**♡* **Hirak Sarkar**♢ **Ashiqur R. KhudaBukhsh**♠
♣Maulana Abul Kalam Azad University of Technology
♡Independent Researcher
♢University of Maryland
♠Carnegie Mellon University
rupaksarkar.cs@gmail.com, sayantan.mahinder@gmail.com,
hsarkar@cs.umd.edu, akhudabu@cs.cmu.edu

Abstract

Attribution of natural disasters/collective misfortune is a widely-studied political science problem. However, such studies typically rely on surveys, expert opinions, or external signals such as voting outcomes. In this paper, we explore the viability of using unstructured, noisy social media data to complement traditional surveys through automatically extracting attribution factors. We present a novel prediction task of *attribution tie detection* of identifying the factors (e.g., poor city planning, exploding population etc.) held responsible for the crisis in a social media document. We focus on the 2019 Chennai water crisis that rapidly escalated into a discussion topic with global importance following alarming water-crisis statistics. On a challenging data set constructed from YouTube comments (72,098 comments posted by 43,859 users on 623 videos relevant to the crisis), we present a neural baseline to identify attribution ties that achieves a reasonable performance (accuracy: 87.34% on attribution detection and 81.37% on attribution resolution). We release the first annotated data set of 2,500 comments in this important domain[1].

1 Introduction

Water crisis is one of the pressing current environmental challenges. More than a billion people do not have access to clean drinking water, and every year nearly two million children die from water borne diseases (Watkins, 2006). One-third of the world's most extensive groundwater systems are under severe stress (Richey et al., 2015). The forecasts look even more grim; nearly two-thirds of the world population could be water stressed by 2025 (Seckler et al., 1999). While the crisis has reached an alarming level far and wide, India is listed as one of the major at-risk countries (Rost et al., 2008). In June 2019, the longstanding Chennai water crisis (WashingtonPost, 2019) escalated into an international talking point, revealing alarming statistics of the water crisis in India looming in near future. In this context, we define a new task of inferring *attribution ties* through large scale analysis of relevant social media discussions. Our main contributions in this paper are the following.

Social: Apportioning attribution for a collective crisis or misfortune still remains a challenge in social science, despite the presence of a large body of political science literature on retrospective voting (see, e.g., Ferejohn 1986; Peffley 1984) or psychological literature on attribution (see, e.g., Shaver 2012). Prior social science literature (Griffin et al., 2008) primarily relies on traditional surveys for attribution analysis. Unlike traditional surveys, social media analyses are vastly cheaper, have faster turnaround time, can be conducted at different spatiotemporal granularities and aggregate a larger number of opinions than traditional surveys can usually afford. For instance, the most-recent PEW survey (Pew) focused on India was conducted in 2018 on only 2,521 users. In contrast, our data set consists of comments from 43,859 users.

Data set on crisis attribution: To the best of our knowledge, we present the first large scale social media analysis of the Chennai water crisis via a substantial corpus of 72,098 YouTube comments posted by 43,859 users on 623 relevant videos. Our choice of YouTube is informed by

[*] Rupak Sarkar and Sayantan Mahinder are equal-contribution first authors. Ashiqur R. KhudaBukhsh is the corresponding author.

[1] Code and data are publicly available at https://www.cs.cmu.edu/~akhudabu/WaterCrisis.html.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1402–1412,
November 16–20, 2020. ©2020 Association for Computational Linguistics

(1) its global reach; (2) its popularity in the Indian subcontinent (HindustanTimes, 2019); and (3) prior literature of analyzing globally important events (Palakodety et al., 2020a,c; Cinelli et al., 2020). We not only analyze and present the nuances of social media conversation in Indian subcontinental English, but we also release the first annotated data set on this important domain.

NLP task and model: Our main machine learning contributions are a new task of detecting *attribution ties* from unstructured web data and baselines that automatically detect them. Table 1 lists a few example comments from our data set. We argue that the task of attribution ties detection is a challenging NLP task that requires subtle understanding of language constructs. Consider the following example: '*stop have **9 kids family***'. In this comment, a growing population is attributed as the possible cause of the water crisis. While there is no surface level text match with the term *'population'*, humans can still infer it from the semantic equivalence of *population'* and *'9 kids family'*. As there can be many equivalent ways of expressing attributions, a semantic understanding of the language is necessary for the task. Moreover, although necessary, establishing semantic equivalence is not sufficient for attribution detection. Consider another example: '*can't feed **9 kids family***'. In this example, we again see that the same phrase *'9 kids family'* is present, yet the comment is not attributing to *'population'* for the water crisis. Hence, to correctly detect an attribution tie, we also need to understand the context in which an attribution factor is mentioned. Finally, scarcity of labelled data and the informal nature of conversation in social media pose additional challenges. We present a spectrum of model architectures with increasing sophistication that encode these topical and contextual information to detect attribution ties from a user comment. We use pre-trained language model (LM) to leverage transfer learning and overcome the challenge of paucity of labelled data. Furthermore, we demonstrate that fine-tuned, pre-trained LMs on Indian social media data have the ability to generalize and tackle the quirks of English written by non-native speakers. We find that applying such models improves the performance on the attribution detection task, even though the underlying LM was tuned on a data set that primarily focused on a topic (2019 Indian General Election) (Palakodety et al., 2020b) different from ours.

Attribution factor	Comment
Overpopulation	people need to stop having kids otherwise this lack of good water problem will spread
Climate change	coastline cities like mumbai and chennai is going to sink under water after sea rise due to global warming while we fight for water
Deforestation	plant trees dumb ass trees will hold water as well as soil you have no trees at all that is why you have not water
Contamination | Public water wastage	from the water truck they poured much of it on the ground they put dead bodies and trash in their own water
Government | Public water wastage	not only government but all the the civilian sorry equally responsible for or the water crisis i live in Delhi and it is not a single day when i have not encounter water wastage

Table 1: Examples of *attribution ties* in our data set. Multiple factors are separated by | .

2 Related Work

Water crisis has received sustained research focus in a diverse set of fields such as food policy research (Hanjra and Qureshi, 2010), earth science (Qin et al., 2007), social science (Foltz, 2002), and water research (Schindler and Donahue, 2006; Narula et al., 2011; von Medeazza, 2006), encompassing a broad range of dimensions including the socio-hydrological, ethical, cultural, and foreign policy aspects of the crisis. Our work relies on these lines of research to compile a list of possible attribution factors (see, Table 3). However, our focus is different as we seek to tackle the NLP challenges associated with analyzing attributions from noisy social media data. Our work shares similar motivations to a recently-reported work on the Flint water crisis (Oz and Bisgin, 2016) that evaluated attributions from a substantial tweet corpus. Our work is different from Oz and Bisgin (2016) for the the following reasons: first, our data set is linguistically more challenging (see Section 3.2) as a vast majority of the content creators are non-native speakers of English, second, we propose a learning problem that automates the detection of attributions while Oz and Bisgin (2016) formed different hypotheses on the nature of the attributions and then accepted or rejected those hypotheses based on randomly sampled data labelled by annotators.

Methodologically, our work is closely related with automatic extraction of blame ties (Liang et al., 2019). Similar to Liang et al. (2019), we seek to extract causal ties (Miwa and Bansal, 2016) between a crisis and different possible factors. However, unlike the present work, Liang et al. (2019) focused on a clean corpus obtained from three major US newspapers. In contrast, we embrace the challenge of detecting *attribution ties* from noisy, social media data which involves the following challenges. First, these discussions are produced in a part of the globe where vast majority of content contributors are non-native speakers of English. Second, social media discussions encompass a diverse set of expressions ranging from stating pure fact or statistics to crude disgust and subtle sarcasm, and attribution topics are often expressed in widely different ways. For example, both the comments, '*we must protect our forests* ***plant more trees***' and '*just rewind and see how many* ***trees have vanished*** *over the years to accommodate more space for buildings and malls*' deemed *deforestation* responsible for water crisis but have different ways of expressing it. In contrast, Liang et al. (2019) dealt with a set of well constructed entities that are easy to detect in a sentence due to their crisp word boundaries.

3 Data Set

3.1 YouTube Video Comments

Using the publicly available YouTube's Search API, we query YouTube with the following search queries: *Chennai water crisis*; and *India water crisis*. For each query, we construct our video set, $\mathcal{V}$, by adding 350 recommended videos. Upon removal of duplicate videos and videos without a single comment, $\mathcal{V}$ is pruned to contain 623 unique videos. For each video in $\mathcal{V}$, we extract posted comments using the publicly available YouTube Data API. Our overall comment data set, $\mathcal{D}_{all}$, consists of 72,098 comments.

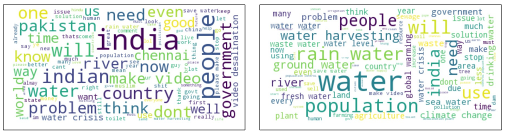
Since India is a country with vast linguistic diversity, a language identification technique is required to extract comments written in English. We use a recently-proposed language identification method (Palakodety et al., 2020a) that has been successfully used for both document and token level language identification (KhudaBukhsh et al., 2020) and other multilingual settings (Palakodety et al., 2020b) and extracted comments written in English. Our filtered set of English comments, $\mathcal{D}$, consists of 41,791 comments.

3.2 Data Set Challenges

Beyond the typical challenges posed by noisy social media texts, in our case, the vast majority of the contributors are non-native English speakers often employing a telegraphic and colloquial style. We outline some of these challenges with representative examples next. A detailed treatment of this challenge is presented in (Sarkar et al., 2020).

Topic 1 (17.6%)	Topic 2 (16.3%)	Topic 3 (9.6%)	Topic 4 (9.5%)	Topic 5 (8.2%)
water	india	change	india	muslim
save	country	climate	pakistan	indian
need	population	global	river	india
drink	people	human	china	religion
river	indian	nature	kashmir	hindu
waste	problem	animal	shit	like

Table 2: Most relevant tokens for five major topics discovered in our data set using Blei et al. (2003).



(a) $\mathcal{D}$ (b) $\mathcal{D}_{pruned}$

Figure 1: Word cloud visualizations of $\mathcal{D}$ and $\mathcal{D}_{pruned}$.

Spelling errors: We notice a considerable amount of phonetic spelling errors (e.g., '*check the* ***expedinjar*** *level in India and other countries*' originally intended to express **expenditure**).

Out of vocabulary (OOV) words: Several comments use contraction (e.g., '***plz*** *make* ***vdo*** *in rainwater harvesting*'), hence generating OOV words. Our data set has only 28.9% intersection of words with GloVe (Pennington et al., 2014) vocabulary.

Grammatical errors: Several comments suffer from grammatical disfluencies (e.g., '*this not happen everyear because of heat wave in south india this happen*') making our analysis challenging.

3.3 Topical Focus

To present a broad overview of the topics, Table 2 summarizes our topic modeling results using LDA (Blei et al., 2003). As shown in Table 2, the main topics of discussion relevant to the crisis involve call to save water (topic 1), overpopulation as a major problem (topic 2), and climate change (topic 3). A considerable fraction of overall discussion is focused on peripheral topics unrelated to the water crisis (topic 4 and topic 5). For example, the presence of topics surrounding India and Pakistan is not surprising since the Pulwama terror attack in Kashmir (Feb, 2019) and an ensuing India-Pakistan conflict was a major contemporaneous sociopolitical issue (BBC).

3.4 Data Pruning

Since the peripheral discussions unrelated to water crisis are not meaningful for our current analysis, in order to reduce annotation cost, we use an embedding-based method to first filter in comments

topographical disadvantage, weather, climate change, global warming, industrial development, petroleum industry, water intensive industries, oil sands development expansion of urban areas, conversion of lands for human usage, urban waste, corruption, mismanagement, contamination, industrial wastewater, industrial waste draining, cyanobacteria, bacteria, overpopulation, population shift, excessive demand, irresponsible irrigation, water intensive irrigation, irrigation water demand, irrigated agriculture, water intensive agriculture, inefficient irrigation, water withdrawals, irresponsible water pumping, public water wastage, excessive usage, indifference of policy makers, lack of funding , funding cuts, lack of study, loss of water bodies, depletion of ground water, permanent removal from water cycle groundwater exploitation, strain on natural resources, deforestation, nutrient loss in soil, eutrophication, drought, flood, damming, impoundment, human activity, water intensive protein rich diet, consumption by livestock, inefficient distribution system

Table 3: List of factors obtained from existing water crisis literature.

more likely to be relevant to the water crisis.

First, we consult relevant research conducted by the water research, urban planning, political science and environmental science communities, and ground our analysis through constructing a list of potential factors scientists typically identify as possible reasons for water scarcity. Our list (presented in Table 3) is based on literature (1) focusing on the global water crisis; (2) targeted analysis on a wide range of geographic regions; and (3) the specific water crisis in Chennai and broadly in India (Schindler and Donahue, 2006; Hanjra and Qureshi, 2010; Qin et al., 2007; Foltz, 2002; Marshall, 2011; Rodell et al., 2009; Narula et al., 2011; von Medeazza, 2006).

Since several factors listed in Table 3 are semantically close, we define 21 broad attribution categories listed in Table 4. We acknowledge that several other reasonable and logical partitions of these attribution categories are possible.

While the list presented in Table 3 is comprehensive covering a broad range of geographical regions, given India's multi-layered socio-political diversity, some of the attribution factors may not be present in the compiled list. In such cases, we instructed the annotators to describe the category in a simple English phrase of not more than four words. For instance, religion was a category discovered by our annotators; a small fraction of comments blamed specific religions for overpopulation and contaminating the Ganges. Similarly, (lack of) desalination facilities was identified as another factor.

Let $\mathcal{F}$ denote the set of factors presented in Table 3. Let a comment d be represented as a sequence of sentences $s_1, \ldots, s_n$. For each s_i, we compute the embedding-based cosine similarity between $\langle s_i, f \rangle$, $f \in \mathcal{F}$ (denoted as *Cosine*$(\langle s_i, f \rangle)$). We use 300 dimensional GloVe (Pennington et al., 2014) embeddings in this step. While calculating the embedding of a sentence, we removed stopwords and OOV words and computed a tf-idf weighted mean of the remaining words. For a given comment, attribution factor pair, $\langle d, f \rangle$, the similarity score, *sim*$(\langle d, f \rangle)$ is defined as *sim*$(\langle d, f \rangle)$ = $\max_i$ (*Cosine*$(\langle s_i, f \rangle)$)) We removed all the $\langle d, f \rangle$ pairs for which - either *sim*$(\langle d, f \rangle)$ is less than 0.7 or f do not fall in the top 20 percentile of the nearest comments of any attribution factor.

Our pruned comment set, $\mathcal{D}_{pruned}$, consists of 2,282 comments (9,004 sentences). A word cloud visualization (see, Figure 1) reveals that our pruning method lends more prominence to water specific tokens than tokens unrelated to the crisis (e.g., Pakistan). We randomly sampled 1,500 comments from $\mathcal{D}_{pruned}$ (6,135 sentences), and 1,000 comments from $\mathcal{D}$ (3,284 sentences) for annotation. The percentages of comments having at least one attribution from $\mathcal{D}$ and $\mathcal{D}_{pruned}$ are 24.30% and 73.87%, respectively (i.e., embedding-based pruning yields more positives than random sampling).

Combining the samples from $\mathcal{D}$ and $\mathcal{D}_{pruned}$, we obtain our final data set of 2,500 annotated comments. Since a comment may consist of multiple sentences with different sentence attributing to different factors, our annotators labeled at the granularity of a sentence. After annotation, we obtain 1,351 comments with at least one attribution. We next merge contiguous sentences (from the same comment) with identical label into a single sentence yielding 2,385 positives and 5,837 negatives.

3.5 Characterizing the Annotated Data

Three annotators proficient in Hindi, English and Bengali conducted annotation in two separate phases. In the first phase, the annotators label if a sentence contains an attribution. A high Fleiss' κ measure of this task (0.86) indicates strong interrater agreement. Next, they specify the attribution factor chosen from the list presented in Table 3. For a given instance, a rater is allowed to choose multiple labels if she deems appropriate. Next, disagreements are resolved through a follow-up adjudication process. Following (Pavlick and Kwiatkowski, 2019), we surface any inherent ambiguity/disagreement between annotators in the final set of labels. Even after the adjudication process, if raters fail to resolve (say, $rater_1$ sticks to attribute a and $rater_2$ and $rater_3$ stick to attribute b), we propagate $\{a, b\}$ as the final label (accounting for 2.4% of the non-singleton labels). We find

Broad Category	Sub-categories
Agriculture	agricultural use, water intensive irrigation, inefficient irrigation, water intensive crops
Climate change	climate change, global warming, weather
Corruption	corruption, mismanagement
Damming	damming, impoundments
Deforestation	deforestation, nutrient loss in soil
Desalination	desalination
Government inaction	government inaction, indifference of policy makers, lack of proper funding
Groundwater exploitation	groundwater exploitation, strain on natural resources
Human activity	human activity, water intensive protein rich diet, consumption by livestock
Industrial development	industrial development, petroleum industry, water intensive industries, oil sands development
Lack of awareness	lack of awareness, lack of study
Lack of infrastructure	lack of infrastructure, inefficient distribution system
Lack of harvesting	lack of rainwater harvesting, lack of water preservation
Loss of water bodies	loss of water bodies, loss of water tables
Natural calamities	drought, flood
Overpopulation	overpopulation, excessive demand, population shift
Pollution	pollution, contamination, industrial waste water, industrial draining
Public water wastage	public water wastage, excessive usage
Religion	religion, Hindu caste system, Islam
Water Withdrawals	water withdrawals, irresponsible water pumping
Urbanization	urbanization, expansion of urban areas, land conversion, urban waste

Table 4: 21 broad categories of attribution factors.

that overpopulation, climate change, deforestation, public water wastage, pollution and government inaction are recurrent themes in the discussion.

4 Model Specification

4.1 Attribution Task

Given a set of YouTube comments $\mathcal{D}$ and a set of attributing factors $\mathcal{F}$ as described in Section 3, we aim to learn the underlying *attribution ties* between a comment $d \in \mathcal{D}$ and the set of attributing factors $f \in \mathcal{F}$. A simple way to model this can be posing the task as a multi-class classification over $\mathcal{F}$. We model this instead as learning a probability density function which determines for a tuple $\langle d, f \rangle$, how likely the factor f is attributed in the comment d. This allows us to learn the attribution relationship over a generic set of factors that may not be completely known *a priori*. Given a set of pairs of $\langle d, f \rangle$ labelled as positives (i.e., f is attributed in d), we aim to learn the different ways people express themselves when they attribute f in d. We define, $\mathcal{A}\colon (\mathcal{D}, \mathcal{F}) \mapsto [0, 1]$ as an attribution function that estimates the probability of the attribution relationship given $\langle d, f \rangle$ pair as input.

The task of designing the attribution function poses the following challenges. At a conceptual level, we need to model the specific topical relationship between d and f where the context of f in d is an attribution and not just a simple mention. In addition, the model needs to operate on the type of language used in social media, taking into account the challenges associated with non-native English speakers (mentioned in Section 3.2). At the implementation level, due to scarcity of labelled data, it is not possible to train an end-to-end LM that can capture all these nuances and hence we used pre-trained LMs such as BERT and its fine-tuned variant BERT$_{Indian}$ (Palakodety et al., 2020b). We address the over-fitting problem caused by low volume of labeled data by constraining the model to use a small number of trainable parameters while learning the underlying LM. As mentioned in Section 1, our proposed model aims to capture both the *topical similarity* of f in d and the *context* in which f is used in d. For every word $w_i \in d$ where $i = 1, 2, \ldots, n$ are the indices of each word in the comment, we define $\mathit{Similarity}_i^f$ as a semantic similarity measure between word w_i and attribute f. Furthermore we define, $\mathit{Context}_i^f$ as the measure that word w_i is used to express the context in which f is mentioned as the attributing factor for the crisis. To compute $\mathit{Similarity}_i^f$, we use an idea similar to attention mechanism by (Bahdanau et al., 2015). Specifically, we use cosine similarity between the representations of the attribution factor f and representations of w_i from the LM. We formulate $\mathit{Context}_i^f$ as an inversely correlated function of $\mathit{Similarity}_i^f$, where our intuition is for a positively labeled data-point, every word w_k that doesn't represent f, must capture its context in d.

4.2 Model Architecture

The model architecture is demonstrated in Figure 2. Applying the above intuition for a $\langle d, f \rangle$ pair (a comment-attribute tuple), we first obtain the contextual word embeddings $e(w_i)$ and $e(w_j)$ for words, $w_i \in d$ and $w_j \in f$ respectively, by using an LM such as BERT. As the attribution factors contain only a few words, we set the representation $E(f)$ for the attribution factor f as the mean embedding, $\mathit{mean}_{j \in |f|}\{e(w_j)\}$ (Eq. 1).

We then construct the probability function, $\mathit{Similarity}_i^f$ (Eq. 3), for a factor f and a word w_i, by using the cosine similarity (denoted as c_i in Eq. 2) between $E(f)$ and $e(w_i)$. The *topical similarity*, $E_{\mathit{topic}}(d)$ (Eq. 5) for the entire comment d is represented as a linear combination of the contextual word embeddings, $e(w_i)$ weighted by individual $\mathit{Similarity}_i^f$. Finally, we use $1 - c_i$ as a loose mea-

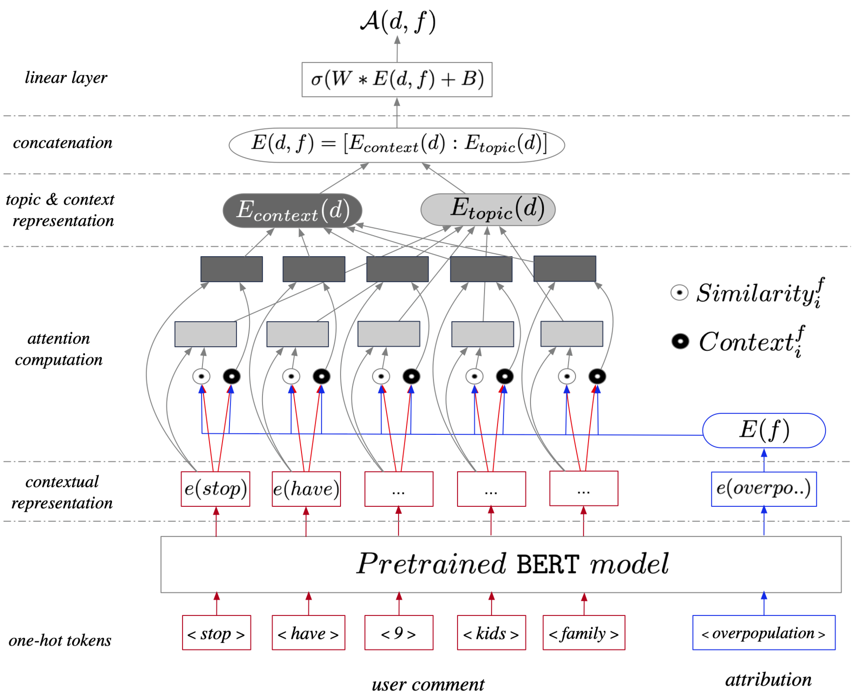


Figure 2: Model architecture

sure of inverse cosine similarity for constructing the probability function, $Context_i^f$ (see Eq. 4) and similarly generate the non-topical contextual representation $E_{context}(d)$ (Eq. 6) for the comment. Note that,

$$E(f) = mean_{j \in |f|}\{e(w_j)\} \quad (1)$$

$$c_i = Cosine(e(w_i),\ E(f)) \quad (2)$$

$$Similarity_i^f = \sigma(\alpha * c_i + \beta) \quad (3)$$

$$Context_i^f = \sigma(\alpha * (1 - c_i) + \beta) \quad (4)$$

$$E_{topic}(d) = \sum_{i \in |d|} Similarity_i^f \ * \ e(w_i) \quad (5)$$

$$E_{context}(d) = \sum_{i \in |d|} Context_i^f \ * \ e(w_i) \quad (6)$$

In Equations 3 and 4, α and β are the hyperparameters and $\sigma(.)$ is the sigmoid function to scale the cosine similarities to $[0, 1]$ range. The concatenation of the $E_{topic}(d)$ and $E_{context}(d)$ is used as the final representation of the $\langle d, f \rangle$ pair:

$$E(d, f) = [E_{topic}(d) : E_{context}(d)] \quad (7)$$

The final representation, $E(d, f)$, is passed through a linear layer with dropouts to model the attribution function $\mathcal{A}$ and is trained with Binary Cross Entropy loss (BCELoss) using binary labels. The linear layer, with learnable parameters W and B, is defined as follows,

$$\mathcal{A}(d, f) = \sigma(W * E(d, f)\ +\ B) \quad (8)$$

5 Experimental Setup

5.1 Model Training

We use an 80:10:10 split to divide the labeled data into training, validation and hold out sets, respectively. Two different pre-trained LM weights are used to bootstrap our model. We first use the basic-BERT ('bert-base-uncased') model by initializing our model with the pre-trained weights obtained from Huggingface's transformer API (Wolf et al., 2019). The weights for the other model BERT$_{Indian}$, were generated by (Palakodety et al., 2020b) where the authors fine-tuned BERT on a large corpus of 2 million comments posted in a 100 day period leading up to the 2019 Indian General Election. The BERT$_{Indian}$ weights boost the performance over BERT weights in our task, as it was trained on linguistic expressions typical to Indian social media. In addition to the BERT variants, we also experiment with few other baseline setups explained in details in Section 5.3.

We use a linear feed-forward layer to convert the language representation vectors to logits. We found adding more layers in the feed-forward network was detrimental towards training; perhaps due to less amount of available training data. For training our models, we used the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1.8\mathrm{e}{-5}$ and a batch size of 4. The hyper-parameters α and β are set to 10 and -5 respectively.

5.2 Performance Measures

We evaluate our models' performance at two different levels. Our original task is to identify if a YouTube comment is attributing to one (or more) attribution factor(s) and deeming it (them) responsible for the water crisis. Since the number of comments without any attribution is significantly higher than the number of comments with attribution, we conduct a fine-grained evaluation of our models' performance in the following way. We divide our prediction task in two sub-tasks: (1) attribution *detection* and (2) attribution *resolution*. While the detection task aims to predict the presence of attribution in a sentence, the resolution task involves correctly identifying the attributed factor. Furthermore, we measure resolution at a conservative top-1 as well as a relaxed top-3 setting to analyze the model's performance with near-similar attributes.

For the detection task, we apply a threshold to determine if the sentence has any attribution at all from a provided set of attribution factors. The threshold value is tuned on the validation set. Hence, the detection task reduces to evaluating the condition $\mathcal{A}(d, f') \geq T$, where $f' = \arg\max_{f \in \mathcal{F}} \mathcal{A}(d, f)$. For resolution, we choose the best attribution factor f' for the conservative top-1 setting. For the relaxed top-3 setting, we order the list of attribution factors $f \in \mathcal{F}$ by the corresponding score $\mathcal{A}(d, f)$ and pick the top three candidates from the list.

In presence of multiple attribution factors, we use a set membership test to assign a binary outcome. Let $\mathcal{F}_{true} \subseteq \mathcal{F}$ denote the ground truth set of attribution factors for a comment d, and our classifier predicts attribution factors $\mathcal{F}_{predicted}$ for d. The binary outcome of the prediction task is $\mathbb{I}(F_{predicted} \cap \mathcal{F}_{true} \neq \emptyset)$, where $\mathbb{I}$ is the indicator function denoting success or failure for the resolution task.

5.3 Baselines and Ablation Study

We now describe our baseline (denoted by $\mathcal{M}_{\texttt{GloVe}}$) and models. We start with a simple model and add sophisticated techniques in subsequent iterations.

- **Word embedding** ($\mathcal{M}_{\texttt{GloVe}}$): We use a GloVe (Pennington et al., 2014) embedding-based similarity measure to establish our baseline. The baseline is inspired by the observation presented in Arora et al. (2017) and emphasizes on the intuition that weighted word embeddings produce high quality sentence representations. We use an idf (inverse document frequency) weighted sum of GloVe word embeddings for all the words in a sentence to compute the sentence representation. Next, we use the same method for the attribution factors to get the attribution representation. A cosine similarity between the sentence and attribution representations is used to determine this baseline with no task-specific training.
- **Classification over BERT** ($\mathcal{M}^{simple}_{\texttt{BERT}}$): $\mathcal{M}^{simple}_{\texttt{BERT}}$ uses an LM based classification technique where we build a linear classifier on top of the pre-trained BERT. We take the mean contextual embeddings, $E(d) = mean_{i \in |d|}\{e(w_i)\}$ for the words $w_i \in d$ as the sentence representation and, $E(f) = mean_{j \in |f|}\{e(w_j)\}$ for the words $w_j \in f$ as the attribution representation. A linear layer is trained over the concatenated vector $[\ E(d) : E(f)\ ]$ to learn the attribution relationship between comment d and factor f. While training, both the new parameters from the linear layer and the underlying BERT parameters are learned. We notice that freezing the BERT parameters to train only the top linear layers yields inferior results for all the LM based setups.
- **Topical similarity model** ($\mathcal{M}^{topic}_{\texttt{BERT}}$): In this model, we only use the topic similarity to create the topical representation of d as $E_{topic}(d)$ (Eq. 5). The factor f is represented by $E(f)$ (Eq. 1). The concatenation of the two $[\ E_{topic}(d) : E(f)\ ]$ is passed to the linear layer and is jointly trained with the language model parameters.
- **Final architecture** ($\mathcal{M}^{final}_{\texttt{BERT}_{Indian}}$): This model uses both $E_{topic}(d)$ (Eq. 5) and $E_{context}(d)$ (Eq. 6) as described in Section 4. We find that the introduction of contextual representation $E_{context}(d)$ over the previous setup performs better.
- **Switching to BERT$_{Indian}$** ($\mathcal{M}^{final}_{\texttt{BERT}_{Indian}}$): The architecture of this setup is identical to $\mathcal{M}^{final}_{\texttt{BERT}}$ with the sole modification being the use of BERT$_{Indian}$ (described in 5.1) instead of BERT.

6 Results

We summarize the performance of our baseline and models in Table 5. Following standard practice in evaluating performance on data sets with class im-

Model	*Metric*	*Detection*	*Resolution*	*Resolution + top 3*
$\mathcal{M}^{final}_{\texttt{BERT}_{Indian}}$	P	75.88	**70.14**	74.10
	R	81.99	61.22	74.57
	F1	**78.81**	**65.38**	**74.34**
	Acc	87.34	**81.37**	**85.20**
$\mathcal{M}^{final}_{\texttt{BERT}}$	P	66.92	59.17	64.53
	R	**92.58**	**66.31**	**83.26**
	F1	77.68	62.54	72.71
	Acc	86.42	79.42	84.47
$\mathcal{M}^{topic}_{\texttt{BERT}}$	P	**81.26**	67.92	**77.62**
	R	70.76	34.53	56.56
	F1	75.65	45.78	65.44
	Acc	86.06	77.54	82.04
$\mathcal{M}^{simple}_{\texttt{BERT}}$	P	74.52	22.54	52.14
	R	83.05	8.26	30.93
	F1	78.55	12.09	38.83
	Acc	**88.07**	68.28	74.19
$\mathcal{M}_{\texttt{GloVe}}$	P	38.30	7.45	15.71
	R	86.95	11.28	26.10
	F1	53.18	8.98	19.61
	Acc	57.62	36.77	40.85

Table 5: Performance comparison of our models and baselines. For a given task and a performance measure, the best model's performance is highlighted in bold. Precision, recall and accuracy are denoted by **P**, **R**, and **Acc**, respectively.

public water wastage	everyone forgot within 2 or 3 month later again forget to save and waste water.
lack of harvesting	last year chennai received crazy rains all that water went in drain if we had harvested it and let it replenish ground water borewells would not have run dry this year
deforestation	we cut trees to build flat malls multi stored buildings
government inaction	discorperted i know where you are coming from but do not blame the farmers i think it is more of a governmental problem but farmers should not be in the reap where you sow
contamination	stop using chemical soaps and liquids so that drainage is not harsh for environment human waste and kitchen waste need to be decomposed in each home
overpopulation | deforestation | pollution	the basic reason is population for everything cause this planet had a limit to hold people and to add more we are doing deforestation polluting our rivers air pollution and wasting water. . .

Table 6: Example instances that our classifier correctly resolved.

balance, we focus on precision, recall and F1 score as performance metrics instead of accuracy. We observe that, on the detection task, all the BERT based models perform similarly. However, on the resolution task, the F1 score substantially improves as we keep adding sophistication to our model architecture.

Since many of the attribution factors are semantically close (e.g., loss of water bodies, water withdrawal), we also consider a relaxed resolution criterion where a resolution is evaluated as correct if the models' top three predictions have an overlap with the ground truth as described in Section 5.2.

climate change | no attribution	there is no proof of climate change droughts and floods are all natural phenomenon they have happened before there were humans also
public water wastage | human activity	the best way to save water is to stop consuming animal products so much of our precious water is used for animal agriculture
government inaction | urbanization	urban people are the reason for water shortage
overpopulation | no attribution	it has nothing to do with population control
human activity | government inaction	otherwise all our development is a waste if the people are being eliminated by carcinogens created due our irresponsible administration

Table 7: Examples of misclassified instances. Misclassified attribution factor is marked with red, ground truth is marked with blue.

6.1 Error Analysis

We now focus on some of the specific examples from both the correctly classified and misclassified sentence-attribution pairs to summarize the strengths and shortcomings of our models. As shown in Table 6, $\mathcal{M}^{final}_{\texttt{BERT}_{Indian}}$ was able to correctly identify attributions in sentences even in the presence of certain degree of grammar disfluency and an absence of the exact attribution factor per se. For instance, our model correctly resolved '***we cut trees*** *to build flat malls multi stored buildings*' to *deforestation* even though the specific root terms of the attribution factor is not present in the sentence. Our model is able to identify the attribution factors correctly, even when the comments are longer with complex discourse structure and grammatical errors. For example, '*discorperted i know where you are coming from but do not blame the farmers i think it is more of a* ***governmental problem*** *but farmers should not be in the reap where you sow*' was correctly attributed to *government inaction*. Furthermore, when multiple attributions are present, for example in comments like - *'the basic reason is* ***population*** *for everything cause this planet had a limit to hold people and to add more we are doing* ***deforestation*** *polluting our rivers air pollution and wasting water. . . '*, our model is able to correctly predict all three attributions at top three with high confidence.

We also notice few failures that can be attributed to shortcomings of BERT-like language models (Table 7). For example, our model predicts *'climate change'* is an attributing factor in the comment *'there is* ***no proof of climate change*** *droughts and floods are all natural phenomenon they have happened before there were humans also'* with high confidence. Our model fails to understand the negation as well as the context; perhaps due to a welldocumented limitation of BERT's inability in handling negation (Kassner and Schütze, 2019).

Finally, we observe cases where the model fails to attribute the labelled factor as its top prediction, but the top three predictions feature the labelled factor. For example, '*its their fault look how they waste the water they poured half of it on the ground while drinking it and other things from the water truck they poured much of it on the ground*' was attributed to *pollution* as top factor but the labelled factor *public water wastage* was scored second highest by the model.

7 Discussion

• **Unseen attribution factor**: Our model can generalize to unseen attributions factors. For instance, with a new dummy attribution factor *pandemic* and input sentence *'this flu caused the water crisis'*, our model is able to predict *pandemic* with the highest probability. This merits a deeper exploration with a holdout attribution set we aim to investigate in future.

• **Flint water crisis**: We were curious to know how our model performs in the wild on a data set of a different water crisis. To this end, we zero in on the Flint water crisis, another major water crisis happening in a completely different part of the globe with predominantly different sets of attribution factors. On a data set of 5,000 comments randomly sampled from 503 YouTube relevant to the Flint water crisis (Butler et al., 2016), our model predicts *government inaction*, *pollution* (subsumes contamination according to Table 4), and *corruption*. A human inspection of randomly sampled 200 comments aligns with out classifier's predictions. Table 8 presents a random sample of example comments detected by our classifier.

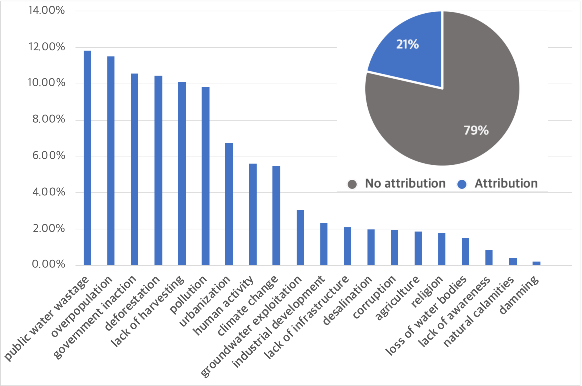


Figure 3: Distribution of number of comments detected by $\mathcal{M}^{final}_{\texttt{BERT}_{Indian}}$ model on 40k comments.

• **The big picture**: We finally run our classifier on our initial data set of 40K English comments to obtain a bigger picture. As shown in Figure 3, we find that nearly 80% of the discussions in our corpus do not contain any attributions. This aligns with our previous annotation experiment that yielded 24.3% positives from randomly sampled comments. Among the attributed comments, we find *public water wastage*, *pollution*, and *overpopulation* are considered as primary causes for this crisis. A human inspection of randomly sampled 200 comments aligns with the classifier predictions. These insights, along with sample comments from the detected attributions, may provide a holistic view of people's opinion around the topic. In the expanding reach of social media, we thus (1) present a new approach to collect aggregated opinions on crisis attribution and complement surveys; (2) focusing on one of the most important crises of the future: water; and (3) release an annotated data set on this important domain.

government inaction	wow that is insane i feel so bad for the people of flint how has the governor kept his job so many people should be punished for this
pollution	the land is poisoned sitting around and wishing for the magical government to fix it is what children do either install water filtering stations like in arizona or move
corruption	rick snyder is a corrupt lying sociopath cutting people off from bottles of clean water is just incredibly cruel they need to vote him out of office

Table 8: Random sample of comments detected as positives from our Flint data set.

References

Kashmir attack: Tracing the path that led to pulwama. https://www.bbc.com/news/world-asia-india-47302467. Online; accessed 29-May-2020.

Pew research center. https://www.pewresearch.org/global/2019/03/25/a-sampling-of-public-opinion-in-india/. Online; accessed 16-Aug-2019.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Blei, Ng, and Jordan. 2003. Latent dirichlet allocation. *JMLR*, (3):993–1022.

Lindsey J Butler, Madeleine K Scammell, and Eugene B Benson. 2016. The flint, michigan, water crisis: a case study in regulatory failure and environmental injustice. *Environmental Justice*, 9(4):93–97.

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.

John Ferejohn. 1986. Incumbent performance and electoral control. *Public choice*, 50(1):5–25.

Richard C Foltz. 2002. Iran's water crisis: cultural, political, and ethical dimensions. *Journal of agricultural and environmental ethics*, 15(4):357–380.

Robert J Griffin, Zheng Yang, Ellen Ter Huurne, Francesca Boerner, Sherry Ortiz, and Sharon Dunwoody. 2008. After the flood: Anger, attribution, and the seeking of information. *Science Communication*, 29(3):285–315.

Munir A Hanjra and M Ejaz Qureshi. 2010. Global water crisis and future food security in an era of climate change. *Food policy*, 35(5):365–377.

HindustanTimes. 2019. Youtube now has 265 million users in india. Online; accessed 20-April-2020.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Shuailong Liang, Olivia Nicol, and Yue Zhang. 2019. Who blames whom in a crisis? detecting blame ties from news articles using neural networks. In *Proceedings of the AAAI*, volume 33, pages 655–662.

Samantha Marshall. 2011. The water crisis in kenya: Causes, effects and solutions. *Global Majority E-Journal*, 2(1):31–45.

Gregor Meerganz von Medeazza. 2006. Desalination in chennai: What about the poor and the environment? *Economic and Political Weekly*, 41(11):949–952.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Kapil Kumar Narula, Ram Fishman, Vijay Modi, and Lakis Polycarpou. 2011. Addressing the water crisis in gujarat, india.

Talha Oz and Halil Bisgin. 2016. Attribution of responsibility and blame regarding a man-made disaster: #flintwatercrisis. *CoRR*, abs/1610.03480.

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. Mining insights from large-scale corpora using fine-tuned language models. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1890–1897. IOS Press.

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020c. 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.

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

Mark Peffley. 1984. The voter as juror: Attributing responsibility for economic conditions. *Political Behavior*, 6(3):275–294.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

BQ Qin, XD Wang, XM Tang, Sheng Feng, YL Zhang, et al. 2007. Drinking water crisis caused by eutrophication and cyanobacterial bloom in lake taihu: cause and measurement. *Advances in Earth Science*, 22(9):896–906.

Alexandra S Richey, Brian F Thomas, Min-Hui Lo, John T Reager, James S Famiglietti, Katalyn Voss, Sean Swenson, and Matthew Rodell. 2015. Quantifying renewable groundwater stress with grace. *Water resources research*, 51(7):5217–5238.

Matthew Rodell, Isabella Velicogna, and James S Famiglietti. 2009. Satellite-based estimates of groundwater depletion in india. *Nature*, 460(7258):999.

Stefanie Rost, Dieter Gerten, Alberte Bondeau, Wolfgang Lucht, Janine Rohwer, and Sibyll Schaphoff.

2008. Agricultural green and blue water consumption and its influence on the global water system. *Water Resources Research*, 44(9).

Rupak Sarkar, Sayantan Mahinder, and Ashiqur R. KhudaBukhsh. 2020. The non-native speaker aspect: *Indian English* in social media. In *Proceedings of the 6th Workshop on Noisy User-generated Text (W-NUT 2020)*, page To appear. Association for Computational Linguistics.

David W Schindler and William F Donahue. 2006. An impending water crisis in canada's western prairie provinces. *Proceedings of the National Academy of Sciences*, 103(19):7210–7216.

David Seckler, Randolph Barker, and Upali Amarasinghe. 1999. Water scarcity in the twenty-first century. *International Journal of Water Resources Development*, 15(1-2):29–42.

Kelly G Shaver. 2012. *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer Science & Business Media.

WashingtonPost. 2019. As a major indian city runs out of water, 9 million people pray for rain. Online; accessed 20-April-2020.

Kevin Watkins. 2006. Human development report 2006-beyond scarcity: Power, poverty and the global water crisis. *UNDP Human Development Reports (2006)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.