Voice for the Voiceless: Active Sampling to Detect Comments Supporting the Rohingyas

Shriphani Palakodety,[1*] **Ashiqur R. KhudaBukhsh,**[2*] **Jaime G. Carbonell**[2]
[1]Onai, 7280 Blue Hill Dr., Suite 10, San Jose, CA 95129
[2]School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, PA 15213
spalakod@onai.com, {akhudabu, jgc}@cs.cmu.edu

Abstract

The Rohingya refugee crisis is one of the biggest humanitarian crises of modern times with more than 700,000 Rohingyas rendered homeless according to the United Nations High Commissioner for Refugees. While it has received sustained press attention globally, no comprehensive research has been performed on social media pertaining to this large evolving crisis. In this work, we construct a substantial corpus of YouTube video comments (263,482 comments from 113,250 users in 5,153 relevant videos) with an aim to analyze the possible role of AI in helping a marginalized community. Using a novel combination of multiple Active Learning strategies and a novel active sampling strategy based on nearest-neighbors in the comment-embedding space, we construct a classifier that can detect comments defending the Rohingyas among larger numbers of disparaging and neutral ones. We advocate that beyond the burgeoning field of hate speech detection, automatic detection of *help speech* can lend voice to the voiceless people and make the internet safer for marginalized communities.

Introduction

On 25th August, 2017 extreme violence was allegedly perpetrated against the Rohingya community in Rakhine state, Myanmar (Thompson 2018). Since then, more than 740,000 Rohingyas (Human Rights Watch 2019) have fled Rakhine state to escape persecution. The Myanmar military's alleged large-scale campaign of ethnic cleansing (Beyrer and Kamarulzaman 2017) has led to one of the fastest-growing refugee crises in the 21st century. However, Myanmar's military and civilian officials have repeatedly denied any claims of atrocities - which are contradicted by extensive evidence (Thompson 2018) and witness accounts indicating widespread genocide or ethnic cleansing.

Modern history has witnessed multiple instances of mass migration of persecuted communities. Our main goal in this paper is not to argue about highly debated issues like Rohingyas' citizenship rights or make politically contentious claims about the Myanmar government, the alleged oppressor's involvement in this crisis. Rather, our goal is to present the first-of-its-kind case-study of what we call a 21st century problem: migrant crisis in the era of ubiquitous internet, where the global audience can weigh in on the matter, shape public opinion about the persecuted community through social media comments, clamor for justice for the oppressed, mobilize help to the community in distress, and perhaps side with the alleged oppressor and paint a picture of distrust, fear and threat about a persecuted minority. In online forums, persecuted communities may have little or no voice in discussions centered around them because (i) much of the discussion occurs in a global language (e.g., English) in which they may have limited proficiency, or (ii) they may have minimal access to internet, or most importantly, (iii) survival is possibly the highest priority demanding a significant chunk of their resources. Online activities disparaging refugee communities may result in serious real world consequences; prior research has even identified a close, causal link between online hate speech and offline violence targeting refugees (Müller and Schwarz 2018).

Contributions:

1. *Domain*: In this paper, via a substantial corpus constructed using comments on YouTube videos (5,153 videos, 263,482 comments posted by 113,250 users) relevant to the Rohingya refugee crisis, we characterize several key aspects of the discourse and show that a medium as powerful as the internet can create an asymmetric discourse where an (allegedly) oppressed minority may have little or no voice to defend themselves from (possibly misinformed) global vitriol. To the best of our knowledge, this is the first AI-focused comprehensive analysis of the Rohingya refugee crisis. In the last decade, besides the Rohingya immigrant crisis, the world has witnessed migrant crises in central America (BBC News 2019b), Venezuela (Goldberg 2019), Italy (BBC News 2019a), and the long-standing Syrian refugee crisis (Office of the United Nations High Commissioner for Refugees 2018) resulting in the displacement of millions of people. We believe our work in characterizing key aspects of the Rohingya migrant crisis will open the gates for similar AI research in this humanitarian domain.

[*]Shriphani Palakodety and Ashiqur R. KhudaBukhsh are equal contribution first authors. Ashiqur R. KhudaBukhsh is the corresponding author.
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2. *Voice for the voiceless*: Existing discourse moderation tools on social media platforms focus on minimizing hate speech through deletion of hostile content and/or flagging belligerent members. We argue that beyond the important field of hate speech detection, automatic identification of user-generated web content that champions the cause of a minority can be equally critical for making the internet a safer, better, and healthier place. For a balanced and nuanced discussion on issues around oppressed minorities, lending greater visibility to pro-minority voice (e.g., through pinning or highlighting content) from a large pool of hostile or ambivalent comments is critical and can be greatly facilitated through automatic methods. To this end, we construct a classifier dubbed the *voice-for-the-voiceless* classifier, that detects content championing the cause of an oppressed minority, in this case the Rohingyas.
3. *Machine Learning*: We propose two new active sampling techniques. Our *voice-for-the-voiceless* classifier is constructed using a novel nearest neighbor active-sampling technique in the comment-embedding space which effectively uncovers rare positives in a negatively-skewed corpus. In addition, we demonstrate that our proposed technique can be extended to another novel nearest neighbor active sampling technique in the user-embedding space to identify sympathetic users and effectively uncover rare positives.

Concise Overview of the Crisis

1982	• In the 135 national ethnic groups listed in new citizenship law, Rohingyas are excluded, effectively rendering them stateless.
Nov 13, 2010	• Aung San Suu Kyi, opposition leader and Nobel peace prize winner, is released from house arrest.
Jun, 2012	• Religious violence erupts in predominantly Rohingya region Rakhine leaving more than 200 dead and close to 150,000 homeless. In the next three years, more than 112,000 flee to Malaysia by boat.
2014	• In Myanmar's first census in three decades, Rohingyas are excluded.
Nov, 2015	• Suu Kyi's party wins in first election post military rule. Rohingyas are not allowed to vote let alone contest.
Oct 9, 2016	• Rohingya insurgent group Arakan Rohingya Salvation Army (ARSA) claims an attack that killed 9 police officers according to the state media. A massive military crackdown ensues triggering an exodus of 87,000 Rohingyas to Bangladesh.
Aug 25, 2017	• State media claims that 12 police officers were killed by ARSA in a coordinated attack on 20 police posts. A large number of Rohingyas flee to Bangladesh as military responds by (allegedly) burning down villages as a part of what they describe as "clearance operations".
Oct 23, 2017	• Since Aug 25, 2017, a continuous stream of Rohingya refugees arrive in Bangladesh with the refugee count reaching more than 600,000.

This timeline illustrates the sequence of events that led to this massive humanitarian crisis (Hunt 2017). As can be seen, arguably, the community experienced a long-standing systemic bias which led to this evolving crisis.

Related Work

Hate speech detection: There is a growing body of literature on analyzing and detecting hate speech in social media such as Facebook (Del Vigna et al. 2017), Twitter (Davidson et al. 2017; Badjatiya et al. 2017), Reddit (Chandrasekharan et al. 2017) and YouTube (Dinakar et al. 2012). While hate speech detection and subsequent intervention (by moderating content or flagging users) are extremely helpful in maintaining a positive web environment, these tools are inadequate in this setting. In our problem, a persecuted community is largely absent in an overwhelmingly negative discussion about them, possibly due to language barriers or simply because they lack sufficient access to the web or social media. Detecting comments that advocate their cause is crucial in representing their views.

Active Learning: We drew inspiration from several existing lines of Active Learning research for constructing our *voice-for-the-voiceless* classifier (Roy and McCallum 2001; Baram, Yaniv, and Luz 2004; Donmez, Carbonell, and Bennett 2007). Since sequentially labeling and retraining models may not be practically feasible, following (Yang and Carbonell 2013), we adopted a batch Active Learning setting to expand our pool of labeled samples. As we shall demonstrate, a large majority of comments are unfavorable to the Rohingyas, making this classification task one with significant class imbalance. Active Learning with class imbalance is a widely studied research problem (see, e.g., (Settles and Craven 2008; Nguyen and Smeulders 2004; Donmez and Carbonell 2008; Tomanek and Hahn 2009)). Our proposed sampling strategy leverages recent advances in language modeling to obtain comment-embeddings (Joulin et al. 2017; Bojanowski et al. 2017) and then mines nearest neighbors in the comment-embedding space to alleviate the class imbalance issue. Our proposed solution can thus can be considered a skew-specialized Active Learning approach. However, unlike (Ertekin 2009), instead of constructing a synthetic sample, our method yields samples from the actual pool of unlabeled data.

In the context of using embeddings to exploit inter-sample similarity for better coverage, our work is related to (Dimovski et al. 2018), however our application domain focuses on a humanitarian challenge involving rare positives whereas (Dimovski et al. 2018) focused on three different data sets (MIT movie, MIT restaurant and ATIS). We present experimental evidence of our technique's robustness in uncovering rare positives starting from seed set examples not even present in the actual corpus. Also, our overall approach melds multiple active learning strategies (e.g., uncertainty sampling, certainty sampling). We considered minority-class certainty sampling since it was found to be useful in reducing class imbalance in short document classification tasks (Attenberg, Melville, and Provost 2010; Sindhwani, Melville, and Lawrence 2009; KhudaBukhsh, Bennett, and White 2015).

Research on migrant crisis: Extensive research on migrant crises including the Rohingya refugee crisis has been performed from a social science perspective (XChange.org 2017; Bhatia et al. 2018; Milton et al. 2017; UN Global Pulse 2017). In what follows, we focus on relevant literature with an AI emphasis. Large-scale social media analysis of the Syrian refugee crisis to explore social and communicative networks in Twitter has been performed in (Lynch, Freelon, and Aday 2014). Using a small set of curated Twitter accounts, community detection has been analyzed in (O'Callaghan et al. 2014).

In terms of domain, our work is most similar to (Chowdhury, Nibir, and Islam 2018) where a classifier to label comments favorable to the Rohingyas' resettlement in Bangladesh was constructed. Our work is different in terms of scale, focus and analysis in the following ways. First, we consider a substantially larger corpus of 263,482 comments on videos retrieved using high-frequency search queries from 19 different countries (listed in Table 1), whereas (Chowdhury, Nibir, and Islam 2018) focuses on 5,000 Bengali tweets generated by Bangladeshis. Presence of multiple countries adds to our linguistic challenges as expression of intent may become more diverse. Second, we provide a comprehensive analysis of the corpus employing topic modeling, user-level analysis to demonstrate underrepresentation of Rohingya sympathizers in the global discussion, and overall sentiment analysis of the corpus using domain-specific sentiment lexicons. Third, our *voice-for-the-voiceless* classifier is more nuanced than merely the sentiment towards settlement in one particular country. Finally, we address a critical challenge of mining positive examples in a rare-class learning problem with a novel approach of nearest-neighbor sampling.

Data Collection

Our data collection process consists of the following steps:

1. We construct a query set, $\mathcal{Q}$ (116 unique queries), by including related queries from Google Trends[1] for the query [Rohingya] from countries listed in Table 1.
2. For each query in $\mathcal{Q}$, we obtained the top 200 YouTube video search results. In addition to these, we performed a targeted crawl focusing on the three months time-duration (July 1, 2017 through September 30, 2017) when the crisis reached its peak. For a given month, for each query in $\mathcal{Q}$, we obtained the top 50 YouTube video search results. Our final consolidated video data set, $\mathcal{V}$, consists of 5,153 unique videos.
3. We used the publicly available YouTube API to crawl the comments for the videos obtaining 263,482 comments posted by 113,250 unique users.
4. Since $\mathcal{Q}$ contains queries from multiple countries where English is not the native tongue, we expected the comment corpus to be a mixed bag of different languages with English being the predominant one. Hence, we required an automated method to identify the English comments for which we used a minimally supervised language filtering approach (Palakodety, KhudaBukhsh, and Carbonell 2019). The size of our English-filtered corpus, denoted as $\mathcal{C}$, is 138,978 comments (i.e., more than 50% of the corpus was written in English). We identified German, Hindi, Bengali, Malay, Urdu, French, and Arabic in our non-English corpus indicating a global presence in the discussion.

[1]https://trends.google.com/trends/?geo=US

Countries considered	Reason for inclusion
Myanmar	Origin country of crisis
Malaysia, Indonesia and Bangladesh	Countries that offered most help
Bangladesh, China, Laos, India and Thailand	Border sharing countries
Germany, Australia, Austria, Canada, Sweden, Norway, Switzerland, Finland, Ireland and Spain	Granted asylum to Syrian refugees
USA	Received 10% of all asylum applications in the OECD countries in 1998-2007 (Haddal 2009).

Table 1: List of countries.

Video statistics: The total number of views of all videos exceeded 15 million. A substantial chunk of the obtained videos come from the month of September, 2017, when the crisis reached its peak. Hence, we believe that our video corpus captured most of the relevant coverage on this issue.

(a) Rohingyas are (b) Rohingyas are not

Figure 1: A word cloud visualization of Rohingyas' perception. Among 992,841 unique bigrams and 2,465,453 unique trigrams, in terms of frequency, [Rohingyas are] and [Rohingyas are not] rank 188^{th} and 358^{th}, respectively.

Analysis

General perception: We first analyze phrases in the comments that match a set of high-frequency text templates. We focus on: [Rohingya are] (or [Rohingyas are]) and the negated variants ([Rohingya are not] and [Rohingyas are not]). The tokens that follow these templates are visualized in Figure 1 demonstrating that the prevalent perception of Rohingyas was largely negative. The templates were chosen by observing the top phrases that occurred in the corpus. A substantial majority often equated them with terrorists and only a tiny fraction of comments stated they were innocent. When the negated versions were

Figure 2: Resettlement debate. Among 2,465,453 unique trigrams, in terms of frequency, [Send them to] ranks 72^{nd}

used (Figure 1b) however, we noticed that innocent became a high-frequency term. Hence, their innocence is questioned by general commenters. Additionally, Figure 1b indicates a debate around the ethnicity of the Rohingyas - several commenters stated that they are neither Burmese nor Bengali, which leads to our next analysis.

Topic 1 (77.5%)	Topic 2 (16.3%)	Topic 3 (3.2%)
people	allah	child
muslims	quran	rapist
rohingya	god	aisha
myanmar	muhammad	puberty
india	islam	mentally
bangladesh	prophet	muhammedans

Table 2: Most relevant tokens for three major topics discovered in the Rohingya corpus.

The resettlement debate: We focused on the template [send them to] to analyze public perception of where they should be resettled and which community or country is responsible for providing assistance and protection. Figure 2 shows that the issue of origin and resettlement is a highly discussed issue in the corpus. Apart from rich Muslim countries, and obvious choices like neighboring countries India and Bangladesh, we were alarmed to notice that [send them to hell] was also a frequently occurring 4-gram in the corpus; among 3,215,489 unique 4-grams, its percentile rank is 99.75.

Aspects of the discussion: We ran the LDA algorithm (Blei, Ng, and Jordan 2003) on our corpus obtaining best results for topic count of 4 (the 4th topic contained code-mixed incoherent tokens). We discover three main themes: (i) a geopolitical discussion centered around India, Bangladesh, and Myanmar - the three geographically significant countries in the area, (ii) a religion-centric discussion with an appeal for help, and (iii) an anti-Islamic cluster primarily consisting of religion-themed slurs. Table 2 contains a list of the most relevant terms per topic. The relevance score is from (Sievert and Shirley 2014).

User level analysis: We were curious to examine: *is it possible to estimate how many Rohingyas engage with videos where global users post highly negative comments about them?* It is not possible to unambiguously identify if a YouTube user is Rohingya or not. However, we identified several Rohingya-focused YouTube channels many of which use the Rohingya language (this language has 1.8 million native speakers as compared to 380 million native speakers of English). Users who predominantly commented on videos hosted by such channels could possibly be Rohingya or Rohingya-sympathizers. Consequently, we divided our set of videos into two mutually exclusive sets: videos that are hosted by Rohingya-focused channels (e.g., Voice of Rohingya) denoted as $\mathcal{V}^{Roh}$ (1,727 videos from 123 channels), the other, denoted as $\mathcal{V}^{other}$ is the complement of $\mathcal{V}^{Roh}$. Videos belonging to $\mathcal{V}^{other}$ (3,426 videos from 1,244 channels) are primarily hosted by News channels (e.g., BBC, Al Jazeera, CNN) and a few individual contributors.

		Politics sub-reddit	India sub-reddit	Induced on $\mathcal{C}$
$\mathcal{C}^{Roh \to other}$	pos	**30.40%**	**25.20%**	**48.57%**
	neg	2.38%	1.85%	2.88%
$\mathcal{C}^{other \to Roh}$	pos	10.01%,	12.45%	18.55%
	neg	32.92%	32.76%	35.30%

Table 3: Percentage of positive and negative comments using lexicons presented in (Hamilton et al. 2016) and a lexicon induced on $\mathcal{C}$.

Of the 113,250 users, 11,326 and 104,973 users commented on $\mathcal{V}^{Roh}$ and $\mathcal{V}^{other}$, respectively. The overlap between the two sets was 3,049 users (Jaccard similarity index 0.0269). Due to disparate size of the two sets, we admit that instead of looking at Jaccard similarity, a more interesting follow up research question could be *who posts more negative comments? Is it the users who are frequent visitors of* $\mathcal{V}^{Roh}$ *but occasionally visit* $\mathcal{V}^{other}$*? Or the other way around?* We focus on the 3,049 users who have commented on at least one video belonging to $\mathcal{V}^{Roh}$ and one video belonging to $\mathcal{V}^{other}$ and define two mutually exclusive user sets: $\mathcal{U}^{Roh \to other}$ (users with more than 80% of comments posted on videos in $\mathcal{V}^{Roh}$) and $\mathcal{U}^{other \to Roh}$ (users with more than 80% of comments posted on videos in $\mathcal{V}^{other}$).

We next obtained English comments made by these two user sets $\mathcal{U}^{Roh \to other}$ (denoted by $\mathcal{C}^{Roh \to other}$) and $\mathcal{U}^{other \to Roh}$ (denoted by $\mathcal{C}^{other \to Roh}$). Lexicon-based sentiment analysis is an established tool for computing sentiment scores (O'Connor et al. 2010). In this scheme, tokens are assigned scores and individual documents' (comments in our case) scores are obtained by combining the constituent token scores. For effective sentiment analysis, obtaining a domain-specific lexicon is crucial (Velikovich et al. 2010). We considered two existing lexicons induced on popular sub-reddits (Hamilton et al. 2016) (politics and India sub-reddits) and a new custom lexicon induced on our corpus using 100-dimensional FastText embeddings (Joulin et al. 2017) and a lexicon inducing algorithm (SENTPROP) (Hamilton et al. 2016). Our test for positive or negative adds

the individual token scores and if the cumulative comment score is greater than 3 (or less than -3), the comment is considered positive (or negative).

As shown in Table 3, across all three lexicons, we found that $\mathcal{U}^{Roh \rightarrow other}$ posted substantially fewer negative comments than positive comments in comparison to $\mathcal{U}^{other \rightarrow Roh}$ where the ratio of positive to negative comments was reversed. Human evaluation on a random sample of 200 comments revealed that a larger share of negative comments posted by $\mathcal{U}^{other \rightarrow Roh}$ were disparaging to Rohingyas, and the small fraction of negative 200 comments posted by $\mathcal{U}^{Roh \rightarrow other}$ were mostly against the Myanmar government's (alleged) atrocities.

Voice for the Voiceless Classifier

We start with pointing out a subtle but important distinction: *Voice-for-the-voiceless speech is not absence of hate speech*. The goals of a hate speech classifier and our voice-for-the-voiceless classifier are different and complement each other. Identifying hateful content for possible moderation certainly has a positive role in making the internet a safer place for a vulnerable community. However, surfacing comments marked as ***not*** *hate speech* does not necessarily lend a voice to the voiceless. For instance, say a user from India respectfully states that India is an over-populated country and does not have enough resources for Rohingyas. This is clearly not hate speech against the Rohingyas, but it is also not voicing the concerns of the voiceless (the Rohingyas).

We next present a definition of *voice-for-the-voiceless* speech and provide examples picked from the corpus or written by us (italicized) to illustrate the point. Understandably, the italicized comments succinctly express a given condition in correct English while examples from the corpus might contain grammatical errors. We note that the definition presented next is specific to the crisis discussed in this paper (the Rohingya migrant crisis). Similar definitions for other crises would require specifying the target and may have additional nuances.

Definition 1: A comment is marked as *voice-for-the-voiceless* speech, if the comment

1. actively seeks to help one or more persons belonging to the (allegedly) oppressed minority (e.g., `[how can we help the Rohingyas]`)
2. urges other people or organizations (such as the UN) to help the (allegedly) oppressed minorities (e.g., *UN should help Rohingyas*)
3. urges other people to come forward and assist or take a humane stance (e.g., `[value the humanity thy r migrating for lives not for luxury]`)
4. advocates for the (allegedly) oppressed community's rights (e.g., *Rohingyas should get Myanmar citizenship*)
5. condemns the atrocities against the (allegedly) oppressed (e.g., *Myanmar government shame on you*)
6. sympathizes with the (allegedly) oppressed community's plight (e.g., `[This just breaks my heart. I wish I could help. All these people commenting about muslims and hindus should be ashamed. The bottom line is these are humans being killed, children being killed. It doesnt matter who started it. It needs to stop!]`)

and a comment is **not** *voice-for-the-voiceless* if it

1. expresses violent intent to a specific entity (including the alleged oppressors) or broad bias against any religious community (e.g., `[Pakistan please nuke Myanmar bhudists]`)
2. calls for aggressive action against the oppressed community (e.g., `deport them all`)
3. demonstrates proverbial whataboutism (e.g., `[what about Yazidis]`)
4. paints a general picture that the community is a threat (e.g.,`[Rohingyas are terrorists]`)
5. shows solidarity with the (alleged) oppressors (e.g., `[well done Myanmar]`)

We mention that that comments neutral to Rohingyas or comments not relevant to this crisis are automatically **not** *voice-for-the-voiceless*.

Active learning with class imbalance: Typically, in Active Learning, a *seed set* of samples is used to construct a classifier which then samples from the unlabeled pool and seeks labels (Settles 2009). For better generalizability of the classifier in the wild, it is often critical that the training set is (i) balanced, i.e., contains sufficient number of examples from both classes (ii) diverse, i.e. captures a wide variety of data points we may encounter in the wild.

We faced the following two research challenges:

- How to obtain a sufficient number of positive comments in a corpus largely disparaging to the Rohingyas?
- How to cover a wide range of aspects of positive (and also negative) comments in our training set so that the classifier performs well in the wild?

Active Learning meets document embeddings: Note that, key phrases (e.g., `send them to`, `deport them all`, `breaks my heart`) may express user intent. However, in a corpus largely filled with negative comments and with a high variance in English proficiency among the contributors, simple mining techniques using exact phrase-level match may not yield sufficient number of positives. In the extreme case, the phrase we are looking for, may not even yield a single exact match. Moreover, the matched comments run the risk of being highly similar to each other and hence may not capture the entire space of varied expressions. Using semantic embeddings to find comments similar to an example positive phrase (or negative phrase) may be effective; however, with a smaller corpus, semantic embeddings may be more prone to inaccuracy (alleviated with a human-in-the-loop in the Active Learning setting). In this work, we meld recent advances in sentence embeddings with Active Learning and propose a novel Active Sampling technique to augment our seed set. The model described in (Pagliardini, Gupta, and Jaggi 2018) is used to obtain a real-valued vector for each comment in the corpus and used to retrieve a comment's nearest neighbors in this embedding space. In conjunction with random sampling, this technique helps us discover a broader, more diverse set of positive examples and helps us combat extreme class imbalance.

Our Active Learning Approach

As illustrated in Figure 3, our approach consists of the following steps:

1. Construct a seed set of positive and negative comments.
2. Expand the seed set by randomly sampling comments from the unseen corpus.
3. Obtain real valued embeddings for the comments, find the nearest neighbors of the seed set and include them in the corpus (new technique presented in this paper).
4. Further expand using minority-class certainty sampling.
5. Perform final expansion using uncertainty sampling.

Seed set: Our seed set (6 positives, 5 negatives) consists of the same set of examples presented in Definition 1.

Random sampling: In order to have better coverage, we randomly sampled 300 comments and labeled them. We obtained 32 positives and 268 negatives, i.e., 10.67% positives. For evaluating our sampling strategies this acts as the baseline. All rounds of manual labeling were performed by two annotators proficient in English. The annotators were presented with the definition, the example seed set, and information on the (alleged) oppressor and the (alleged) oppressed minority. They were first asked to label independently, and then allowed to discuss and resolve the disagreed labels. We obtained strong agreement in every round (lowest Cohen's κ coefficient across all rounds was 0.8766 indicating strong inter-rater agreement).

Nearest-neighbor sampling (NN sampling): For each comment in $\mathcal{C}$, we use a well-known document embedding model from (Pagliardini, Gupta, and Jaggi 2018) to obtain a real-valued embedding. Starting from the seed set, we obtained the seed embeddings and then obtained the comments from the unlabeled corpus whose embeddings were closest to the seed embeddings (i.e. the nearest neighbors). Following (Demszky et al. 2019; Pagliardini, Gupta, and Jaggi 2018) cosine distance was used as the distance metric.

can someone tell me where i can ***help*** charity to them
all the countries should take a stand for these people and force mayanmar government to ***accept them***
No country is too small to ***take on refugees*** and camp them for period of time until the problem is solved by the world leaders making every problem a political issue is just creating dangerous matters for the poor public in some countries animals are cared for more than the humans
sanction myanmar till they understand international law and ***give up ethnic cleansing***

Table 4: Random sample of positive comments obtained using the nearest-neighbor sampling.

Advantages of our technique: First, it allows flexibility while specifying an example comment. Without sufficient knowledge of the corpus, it may be difficult to uncover a rare positive satisfying a particular aspect of the target concept. Some of the examples in our seed set (italicized) did not have an exact match in the actual corpus yet the semantic similarity technique uncovered similar rare positives.

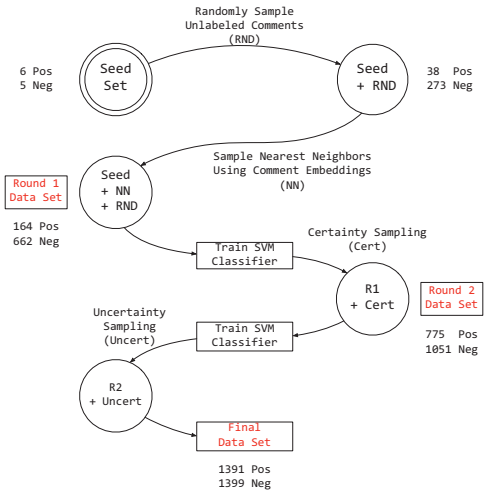

Figure 3: System diagram.

Second, our embedding method employs sub-word information and thus is robust to spelling variations or outright erroneous spellings (e.g., Buddhists was incorrectly spelled as bhudists). In addition, it can handle both short and long comments. In Table 4, we list a few positives we obtained through our technique to highlight its effectiveness. Out of the 300 nearest neighbors obtained from 6 positive seed comments, we obtained 101 unique positives (33.67%) from 292 unique comments. The large number of unique comments indicates that our technique found a diverse set of samples. We obtained more than 3x the number of positives than discovered by random sampling (10.67%)[2]. Our method also uncovered a diverse set of negative comments about the Rohingyas; a representative sample is listed in Table 5 to emphasize why we believed the community required protection from online attacks.

We next trained a Support Vector Machine (SVM) classifier on our consolidated labeled data set with 164 unique positives and 662 unique negatives (we also included the randomly sampled instances) with token unigram, bigram and trigram features.

Certainty sampling. While our *NN sampling* technique effectively uncovered a considerable number of rare positives, the class imbalance was still present with positives merely constituting 19.85% of the labeled data set. We bridged this gap through employing *certainty sampling*, a sampling technique first proposed in (Sindhwani, Melville, and Lawrence 2009; Attenberg, Melville, and Provost 2010). In batch certainty sampling, we pick k (set to 1000) unlabeled samples with highest predicted probability for the minority class. In this step, we closed the gap between the number of positives

[2]We conducted an additional experiment with multiple batches with smaller batch-size to obtain confidence intervals. In this setting, we considered 10 mini-batches of 30 randomly sampled comments (all different from the previously obtained 300 randomly sampled comments) and 10 mini-batches of 30 NN-sampled comments. Across 10 mini-batches, NN-sampling yielded $35 \pm 6.09\%$ positives while random sampling yielded $11 \pm 5.17\%$ positives.

rohingyas are very ***strong in breeding***
rohingya muslims ***are terrorists*** they have been killing buddhist from 1947 onwards ***they deserve*** whatever they are getting
kick them all out ***fuc ing like swines*** and ***changing our demographics***
just ***kill them all*** soon because they are ***terrorists bastards***

Table 5: Random sample of negative comments obtained using the nearest-neighbor sampling.

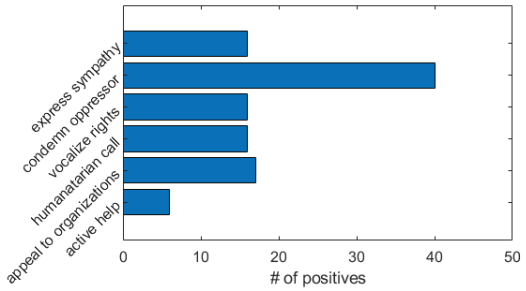

Figure 4: Breakdown of positive comments found in the wild. A single comment can satisfy multiple criteria.

and negatives as we obtained 611 unique positives and 389 unique negatives. We re-train our classifier with our consolidated data set of 775 positives and 1,051 negatives.

Uncertainty sampling. Finally, we used uncertainty sampling to add 1000 more samples where the predicted class probability was close to 0.5. Our final data set consists of 2790 comments with roughly equal numbers of positives and negatives (1,391 positives and 1,399 negatives). Hence, via (i) active learning, (ii) combining multiple existing sampling techniques and (iii) our proposed nearest-neighbor sampling, we succeeded in addressing class imbalance.

Classifier performance. We used a 90/10 train/test split; we trained an SVM classifier (Tong and Koller 2001) with token n-grams as features (with n up to 3) and evaluated the performance on the test set. Since performance can be sensitive to individual test/train splits, we repeated the experiment 100 times on 100 randomly chosen test-train splits. Our intermediate classifiers and final classifier's performance is summarized in Table 7. In a class-imbalanced problem, simply predicting the majority class can yield a high accuracy, F1 score is the more meaningful measure. After each round of labeled data acquisition, we noticed a steady rise in the F1 score with a final performance of $76.50 \pm 2.85\%$. We obtain further improvement by adding comment-embeddings as features as shown in Table 8.

Performance in the wild: Our goal is to identify comments supporting a persecuted minority in the wild. We ran our classifier on the unlabeled corpus (i.e., on comments neither belonging to the train nor test set) and conducted a human evaluation of the top 100 comments predicted as *voice-for-the-voiceless* ranked by confidence. Of the 100 comments, 88% were annotated as positives (vs 10.67% with random sampling) indicating substantial reduction in manual effort to find supportive comments for Rohingyas in the wild. In Figure 4, we present the breakdown of the positives into six broad categories as presented in our definition (active help, appeal to organizations, humanitarian call, vocalize rights, condemn oppressor, express sympathy). We found that our classifier found comments belonging to all broad categories from the wild. In Table 6, we highlight few randomly sampled comments to illustrate two points. First, we draw attention to the bold and italicized comment. We suspect that this comment is written by a Rohingya YouTube user. Broken sentences, grammatical disfluency and a large number of spelling errors indicate how the language barrier may make it difficult for a marginalized community to voice their opinion. Our classifier correctly labeled this comment with high confidence, indicating our approach holds promise in surfacing minority voices. However, since the minority is experiencing (alleged) persecution, it is reasonable to observe measured expression of rational negativity while condemning the (alleged) oppressor. The other italicized comment opens up an interesting philosophical question: ***where should we draw the line?*** For instance, one particular comment supporting the Rohingyas found in the wild used a gendered insult to refer to the Prime Minister of Myanmar which our annotators marked as negative. This underscores the importance of precisely defining the annotation

keep helping these poor innocent people the myanmar government is really kind of like animals not like human beings so thats why they genocide innocent people in myanmar
i am from nepal where buddha was born i have seen buddhist who is so kindful n helpful but i never seen buddhist who murder poor n innocent people i really fell so shameful that they are killing innocent poor people children and old age people they are torturing kid and womens for god shake please stop this violence
thank you so for news today and vi want full human rights in arakan myanmar and stop nvc card and ples vi want myanmar army goverment to the icc kireminal courd justice and vi want full setizenthip in arakan myanmar vi no bangali vi setizenthip in arakan myanmar and myanmar army reped womens rohingya and etnik kilingsing of rohingya and genocide of rohingya and ples vi want hlep from un konsiel and from human rights wohc ples hlep stop genocide of rohingya and humanty in myanmar and thank you so lot god bles you all
its genocide ethnic cleansing brutality reach the level of where words cannot describe its inhuman government of myanmar monk are killing babies and womens into pieces sushi should be punished by court of law

Table 6: Performance in the wild.

Performance measure	Seed set + random sampling + NN in the embedding space	+ Certainty sampling	+ Uncertainty sampling
Precision	$67.17 \pm 9.90\%$	$71.27 \pm 5.23\%$	$73.65 \pm 3.45\%$
Recall	$32.35 \pm 7.65\%$	$72.52 \pm 4.23\%$	$79.39 \pm 3.72\%$
Accuracy	$82.04 \pm 2.34\%$	$75.95 \pm 3.10\%$	$75.38 \pm 2.76\%$
F1 score	$43.02 \pm 7.90\%$	$71.75 \pm 4.32\%$	$76.34 \pm 2.77\%$
AUC	$83.61 \pm 2.88\%$	$83.64 \pm 2.84\%$	$83.67 \pm 2.61\%$

Table 7: *Voice-for-the-voiceless* classifier performance.

Performance measure	SVM (n gram)	SVM (n gram + embedding)
Precision	$73.65 \pm 3.45\%$	**76.49** ± 3.51
Recall	$79.39 \pm 3.72\%$	**80.30** ± 3.73
Accuracy	$75.38 \pm 2.76\%$	**77.71** ± 2.56
F1 score	$76.34 \pm 2.77\%$	**78.28** ± 2.71
AUC	$83.67 \pm 2.61\%$	**85.91** ± 2.32

Table 8: Model improvement.

task. In a similar context, a thorough blue-print is presented in (Olteanu et al. 2018). We conclude this analysis by saying that our classifier holds promise to substantially lessen the burden of moderators to automatically find content supporting a minority, however it may require some further supervision and human judgement to ensure fairness.

Voice-for-the-voiceless community

We conclude our paper with a small exploratory study on the possibility of finding rare positives through user-embeddings. For a given user, we constructed the corresponding user-embedding using the embeddings of all comments posted by the user, normalizing them and finally averaging these normalized embeddings. Our user-focused nearest-neighbor sampling consists of the following steps: (1) Obtain top k positive comments (ranked by predicted class probability) predicted by the *voice-for-the-voiceless* classifier. (2) Next, identify the set of unique users, $\mathcal{U}_{top}$, who posted these comments. (3) Next, for each user in $\mathcal{U}_{top}$, obtain m nearest neighbors in the user-embedding space. (4) Finally, sample comments from the nearest neighbors.

We set both k and m to 10. We obtained 9 unique users who posted the top 10 comments. Of the 90 nearest neighbors, 88 were unique indicating our user-focused nearest neighbor sampling was able to uncover a diverse set of users. We next randomly sampled 300 comments posted by the nearest neighbors. Our hypothesis was if our user embedding-based sampling indeed identifies a set of positive users, the sampled comments will have more positives than the baseline (random sampling fetched 10.67% positives). Our annotators identified 105 positives (35%). Hence, our user-focused sampling performed 3x better than the baseline. Hence, both Active Sampling strategies proposed in this paper substantially outperformed the baseline in finding rare positives supporting a persecuted minority. We conclude with the hope that this work will motivate further AI research in this important humanitarian domain.

References

Attenberg, J.; Melville, P.; and Provost, F. 2010. A unified approach to active dual supervision for labeling features and examples. In *ECML/PKDD*, 40–55. Springer.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 759–760. International World Wide Web Conferences Steering Committee.

Baram, Y.; Yaniv, R. E.; and Luz, K. 2004. Online choice of active learning algorithms. *Journal of Machine Learning Research* 5(Mar):255–291.

BBC News. 2019a. Italy migrant crisis: Government passes tough bill. [Online; accessed 12-May-2019].

BBC News. 2019b. Us-mexico border official says migrant crisis 'at breaking point'. [Online; accessed 12-May-2019].

Beyrer, C., and Kamarulzaman, A. 2017. Ethnic cleansing in myanmar: the rohingya crisis and human rights. *The Lancet* 390(10102):1570–1573.

Bhatia, A.; Mahmud, A.; Fuller, A.; Shin, R.; Rahman, A.; Shatil, T.; Sultana, M.; Morshed, K. M.; Leaning, J.; and Balsari, S. 2018. The rohingya in cox's bazar: When the stateless seek refuge. *Health and human rights* 20(2):105.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3(Jan):993–1022.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW):31.

Chowdhury, H. A.; Nibir, T. A.; and Islam, M. S. 2018. Sentiment analysis of comments on rohingya movement with support vector machine. *arXiv preprint arXiv:1803.08790*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Gentzkow, M.; Shapiro, J.; and Jurafsky, D. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 17th Annual NAACL)*.

Dimovski, M.; Musat, C.; Ilievski, V.; Hossmann, A.; and Baeriswyl, M. 2018. Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings. In

Proceedings of the 27th International Joint Conference on Artificial Intelligence, 4019–4025. AAAI Press.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):18.

Donmez, P., and Carbonell, J. G. 2008. Paired-sampling in density-sensitive active learning.

Donmez, P.; Carbonell, J. G.; and Bennett, P. N. 2007. Dual strategy active learning. In *European Conference on Machine Learning*, 116–127. Springer.

Ertekin, S. 2009. Learning in extreme conditions: Online and active learning with massive, imbalanced and noisy data.

Goldberg, M. L. 2019. Venezuela is a refugee crisis. [Online; accessed 12-May-2019].

Haddal, C. C. 2009. Refugee and asylum-seeker inflows in the united states and other oecd member states. Congressional Research Service, Library of Congress.

Hamilton, W. L.; Clark, K.; Leskovec, J.; and Jurafsky, D. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP*, volume 2016, 595. NIH Public Access.

Human Rights Watch. 2019. Rohingya crisis. [Online; accessed 20-Nov-2019].

Hunt, K. 2017. Rohingya crisis: How we got here. [Online; accessed 12-May-2019].

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th EACL: Volume 2, Short Papers*, 427–431.

KhudaBukhsh, A. R.; Bennett, P. N.; and White, R. W. 2015. Building effective query classifiers: a case study in self-harm intent detection. In *Proceedings of the 24th ACM CIKM conference*, 1735–1738. ACM.

Lynch, M.; Freelon, D.; and Aday, S. 2014. *Syria's socially mediated civil war*. Universitäts-und Landesbibliothek Sachsen-Anhalt.

Milton, A.; Rahman, M.; Hussain, S.; Jindal, C.; Choudhury, S.; Akter, S.; Ferdousi, S.; Mouly, T.; Hall, J.; and Efird, J. 2017. Trapped in statelessness: Rohingya refugees in bangladesh. *International journal of environmental research and public health* 14(8):942.

Müller, K., and Schwarz, C. 2018. Fanning the flames of hate: Social media and hate crime. *Available at SSRN 3082972*.

Nguyen, H. T., and Smeulders, A. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first ICML*, 79. ACM.

O'Callaghan, D.; Prucha, N.; Greene, D.; Conway, M.; Carthy, J.; and Cunningham, P. 2014. Online social media in the syria conflict: Encompassing the extremes and the in-betweens. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 409–416. IEEE Press.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Office of the United Nations High Commissioner for Refugees. 2018. Seven years on: Timeline of the syria crisis. [Online; accessed 12-May-2019].

Olteanu, A.; Castillo, C.; Boy, J.; and Varshney, K. R. 2018. The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018*.

Palakodety, S.; KhudaBukhsh, A. R.; and Carbonell, J. G. 2019. Kashmir: A computational analysis of the voice of peace. *CoRR* abs/1909.12940.

Roy, N., and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *(ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, 441–448.

Settles, B., and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of EMNLP*, 1070–1079.

Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Sievert, C., and Shirley, K. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63–70.

Sindhwani, V.; Melville, P.; and Lawrence, R. D. 2009. Uncertainty sampling and transductive experimental design for active dual supervision. In *Proceedings of the 26th ICML*, 953–960. ACM.

Thompson, N. 2018. Myanmar: Un fact-finding mission releases its full account of massive violations by military in rakhine, kachin and shan statesa. [Online; accessed 12-May-2019].

Tomanek, K., and Hahn, U. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, 105–112. ACM.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *JMLR* 2(Nov):45–66.

UN Global Pulse. 2017. Social media and forced displacement: Big data analytics & machine learning. *UN Global Pulse and UNHCR Innovation Service*.

Velikovich, L.; Blair-Goldensohn, S.; Hannan, K.; and McDonald, R. 2010. The viability of web-derived polarity lexicons. In *NAACL*, 777–785. Association for Computational Linguistics.

XChange.org. 2017. The rohingya survey 2017. [Online; accessed 12-May-2019].

Yang, L., and Carbonell, J. 2013. Buy-in-bulk active learning. In *Advances in neural information processing systems*, 2229–2237.