Hope Speech Detection: A Computational Analysis of the Voice of Peace

Shriphani Palakodety1,2 and Ashiqur R. KhudaBukhsh3,4 and Jaime G. Carbonell4

Abstract. The recent Pulwama terror attack (February 14, 2019, Pulwama, Kashmir) triggered a chain of escalating events between India and Pakistan adding another episode to their 70-year-old dispute over Kashmir. The present era of ubiquitous social media has never seen nuclear powers closer to war. In this paper, we analyze this evolving international crisis via a substantial corpus constructed using comments on YouTube videos (921,235 English comments posted by 392,460 users out of 2.04 million overall comments by 791,289 users on 2,890 videos). Our main contributions in the paper are three-fold. First, we present an observation that polyglot word-embeddings reveal precise and accurate language clusters, and subsequently construct a document language identification technique with negligible annotation requirements. We demonstrate the viability and utility across a variety of data sets involving several low-resource languages. Second, we present an analysis on temporal trends of pro-peace and pro-war intent observing that when tensions between the two nations were at their peak, pro-peace intent in the corpus was at its highest point. Finally, in the context of heated discussions in a politically tense situation where two nations are at the brink of a full-fledged war, we argue the importance of automatic identification of user-generated web content that can diffuse hostility and address this prediction task, dubbed hope-speech detection.

1 INTRODUCTION

"In peace, sons bury their fathers. In war, fathers bury their sons.". – This comment quoting Herodotus, was automatically discovered by our hope speech classifier in our corpus.

On February 14, 2019, a suicide bomber attacked a convoy of vehicles carrying Indian Central Reserve Police Force (CRPF) personnel in Pulwama district, Jammu and Kashmir, resulting in the deaths of 40 CRPF service-personnel and the attacker. A Pakistan-based Islamist militant group claimed responsibility, though Pakistan condemned the attack and denied any connection to it. The Pulwama attack triggered a chain of events where each passing day led to an escalation of tensions between India and Pakistan reaching a peak on the 271/6 of February, 2019. With the two nuclear powers coming precariously close to declaring a full-fledged war, the world witnessed a first-of-its-kind specter of war between nuclear adversaries in the modern era of ubiquitous internet, where a unique war-dialogue took place between the two nations’ civilians on social media.

In this paper, we focus on the discourse that took place in comments posted on YouTube - one of the most popular social media platforms in the Indian sub-continent. We collected a comprehensive data set of comments posted in response to YouTube videos on news coverage of relevant incidents by Indian, Pakistani and global media, and analyzed several important aspects of the dialogue between the two conflicting neighbors in relation to this crisis. To the best of our knowledge, ours is the first large-scale analysis of an evolving international crisis between two nuclear adversaries at the brink of a full-fledged war through the lens of social media. India and Pakistan have a long history of political tension that includes four wars and multiple skirmishes resulting in significant military and civilian casualties [33]. Recent scientific analysis projects 100 million deaths in the event of a full-blown war between these two nuclear powers [19]. As previously presented in [41], social media would play an increasingly important role in understanding and analyzing modern conflicts, and we believe that our work would complement the vast literature of quantitative political science on conflict analysis (see, e.g., [10]).

Contributions: Our contributions are the following:

1. Linguistic: We present a novel language identification technique that requires minimal human annotation based on the observation that polyglot word-embeddings reveal precise and accurate language clusters. Our technique has applications in analysis of social media content in multilingual settings like India, a country with tremendous linguistic diversity (22 major recognized languages) featuring several low-resource languages.

2. Social: Through an extensive polarity phrase lexicon, we analyze the temporal trends of pro-war and pro-peace intent and observe that the pro-peace intent reached its peak when the two nations were closest to declaring a full-fledged war.

3. Hope speech: We propose a novel task: hope-speech detection to automatically detect web content that may play a positive role in diffusing hostility on social media triggered by heightened political tensions during a conflict. Our results indicate that such web content can be automatically identified with considerable accuracy. Solutions to detect hostility-diffusing comments may also find applications in many other contexts. For instance, hostile messages and rumors on platforms like WhatsApp have been used to incite communal violence in the Indian subcontinent in recent times. The severity of the issue prompted the then administration to disable internet access in the regions of unrest to prevent further spread of hateful messages. Beyond a warlike situation, we expect our work to find application in these and other similar settings.

---

1 Ashiqur R. KhudaBukhsh and Shriphani Palakodety are equal contribution first authors. Ashiqur R. KhudaBukhsh is the corresponding author.
2 Onai, USA, email: spalakod@onai.com
3 Carnegie Mellon University, USA, email: akhudabu@cs.cmu.edu
4 Carnegie Mellon University, USA, email:jgc@cs.cmu.edu

---

5 Resources and additional details are available at: https://www.cs.cmu.edu/˜akhudabu/HopeSpeech.html
6 https://www.wired.co.uk/article/whatsapp-web-internet-shutdown-india-turn-off


2 BACKGROUND

Feb 14th  
A suicide bomber kills 40 CRPF personnel at Pulwama, India.

Feb 15th  
A Pakistan-based Islamist militant group, Jaish-e-Mohammad (JEM), claims responsibility. Pakistan condemns the attack and denies any connection to it.

Feb 16th  
India withdraws “most favored nation” status of Pakistan.

Feb 18th  
Nine people, including four Indian soldiers and a policeman are killed in a gun battle in India-controlled Kashmir.

Feb 19th  
Pakistan Prime Minister Imran Khan offers assistance to investigate the Pulwama attack. India refuses the offer citing previous attacks.

Feb 20th  
India halts a bus-service between India-controlled Kashmir and Pakistan-controlled Kashmir.

Feb 23rd  
India begins a two-day crackdown of separatists in Kashmir heightening tensions further.

Feb 26th  
India reports an airstrike against JEM training base at Balakot and reports a large number of terrorists, trainers and senior commanders have been killed. Pakistan denies any such casualty count.

Feb 27th  
As an ominous sign of nuclear threat, Pakistan media reports that Imran Khan chaired a meeting of the National Command Authority, the overseeing body of the country’s nuclear warheads.

Feb 27th  
An Indian Air Force pilot, Abhinandan, is captured by Pakistani armed forces inside Pakistan air space.

Feb 28th  
Pakistan announces that they will release Abhinandan as a peace gesture.

Mar 1st  
Pakistan hands over Wing Commander Abhinandan to India at the Wagah border.

A brief history of the conflict: Kashmir has been a point of contention between India and Pakistan for nearly 70 years. A key factor for continued unrest in South Asia, the Kashmir issue has drawn wide attention from the political science community for decades [19, 33, 3]. The root of this conflict can be traced back to the independence struggle of India and the subsequent partition into India and Pakistan in 1947. Overall, India and Pakistan have gone to full-fledged war four times (1947, 1965, 1971 and 1999) of which, the 1971 war was the goriest (11,000 killed from both sides) which resulted in the largest number of prisoners of war (90,000 POWs) since the Second World War [1]. In the four wars, overall, an estimated 27,650 soldiers were killed and thousands wounded.

Timeline of the most-recent crisis: We outline some of the key events relevant to the most-recent crisis (presented above). We denote five key events: Pulwama terror attack as PULWAMA (Feb 14, 2019), Balakot air strike claimed by Indian Government as an act of retaliation as BALAKOT (Feb 26, 2019), Indian Air Force (IAF) wing commander Abhinandan’s capture by Pakistan (Feb 27, 2019) as IAFPILOT- CAPTURE, Pakistan Government’s subsequent announcement of his release as IAFPILOT-RELEASE (Feb 28, 2019), and Pakistan Government’s handing over of the captured Indian pilot as IAFPILOT-RETURN (Mar 1, 2019).

3 RELATED WORK

Due to our work’s multi-disciplinary nature, throughout the paper, we introduce relevant literature as and when an existing concept is referred. In what follows, we outline a brief description of different lines of research relevant to our paper.
and \( n \). For instance, [Pulwama CNN] is obtained by concatenating the query [Pulwama] and news channel [CNN]. This final set of queries is called \( S_{final} \). Overall, \( S_{final} \) contains 6,210 queries of which 207 queries are obtained from Google Trends.

**Constructing \( V_{relevant} \), a set of relevant videos:** For each query \( q \) in \( S_{final} \), we execute a search using the YouTube search API to retrieving 200 most relevant videos posted during the period of interest. This step yields a result set \( V \) (6,157 unique videos). After removing irrelevant (annotation criterion and details presented in extended version) and unpopular videos (less than 10 comments), we finally obtained \( V_{relevant} \) of a set of 2,890 videos.

**Constructing \( C_{all} \), the overall comments corpus:** For each video \( v \) in \( V_{relevant} \), the YouTube API is used to retrieve all the comments posted during the period of interest. The overall comments corpus, \( C_{all} \), consists of 2,047,851 comments posted by 392,460 users.

**Constructing \( C_{en} \), the English comments corpus:** We next extract English comments using a novel polyglot embedding based method first proposed in this paper (described in the results section). Our English comments corpus, \( C_{en} \), consists of 921,235 English comments.

<table>
<thead>
<tr>
<th>Country</th>
<th>Mention Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>(4,143)</td>
</tr>
<tr>
<td>Pakistan</td>
<td>(2,161)</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>(345)</td>
</tr>
<tr>
<td>Nepal</td>
<td>(234)</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>(85)</td>
</tr>
<tr>
<td>China</td>
<td>(345)</td>
</tr>
<tr>
<td>Russia</td>
<td>(35)</td>
</tr>
<tr>
<td>United States</td>
<td>(234)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>(163)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>(197)</td>
</tr>
<tr>
<td>India</td>
<td>(4,143)</td>
</tr>
<tr>
<td>Pakistan</td>
<td>(2,161)</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>(345)</td>
</tr>
<tr>
<td>Nepal</td>
<td>(234)</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>(85)</td>
</tr>
<tr>
<td>China</td>
<td>(345)</td>
</tr>
<tr>
<td>Russia</td>
<td>(35)</td>
</tr>
<tr>
<td>United States</td>
<td>(234)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>(163)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>(197)</td>
</tr>
</tbody>
</table>

Table 1: Country with mention counts in brackets.

**Investigating coverage:** It is important that the corpus reflects comments from both conflicting countries. We conducted a text template-matching analysis to estimate the origin of the comments posted. We manually inspected the corpus and observed that typically, origin and nationality were expressed through the following phrases: [I'm], [I am], [I am from], [I am an], [I am in], [I am in the], [I am from the], and [love from]. We used these templates and retrieved five tokens following each phrase. Country mentions are extracted from these following tokens and mention frequencies obtained. Using the above heuristics, overall, we assigned nationality to 7,806 users out of 392,460 users (1.99%) accounting for 5.82% of the comments present in the corpus. A log-scale choropleth visualization is shown in Figure 1 and the 10 most mentioned countries are listed in Table 1.

![Figure 1](http://worldpopulationreview.com/countries/united-kingdom-population)

**Figure 1:** Global participation log-scale choropleth. The darker regions indicate larger mention counts (see, Table 1 for actual counts of top 10 countries). Countries in the Indian subcontinent and those with large Indian and Pakistani diaspora populations feature heavily in the discussion.

This analysis illustrates that our corpus contains (i) a balanced participation from both conflicting countries, and (ii) moderate participation from neighboring countries likely to be affected in the event of a war. Interestingly, the plot indicates participation from nations with a significant Indian and Pakistani origin population (USA, United Kingdom, South Africa). We conjecture that modern migration patterns, the nuclear arsenals of India and Pakistan, and the broad global spread of Indian and Pakistani diaspora could be possible reasons for the global attention.

5 RESULTS AND ANALYSIS

5.1 Language identification

Mining a multilingual corpus for insights requires separating out portions of the corpus written in distinct languages. This is a critical step since annotators might be proficient in only a subset of the languages, and the majority of NLP tools are designed for monolingual corpora. We now present an important result to navigate multilingual social media corpora like those generated in the Indian subcontinent.

**Polyglot word-embeddings discover language clusters.** Polyglot word-embeddings are real-valued word-embeddings obtained by training a single model on a multilingual corpus. Polyglot word-embeddings have received attention recently for demonstrating performance improvements across a variety of NLP tasks. While the downstream impact of the embeddings has been explored, in the context of language identification, we perform the first qualitative and quantitative analysis of this embedding space for a variety of Indian and European languages and present the following observations: (i) The word-embedding space is divided into highly-accurate language clusters, (ii) a simple algorithm like \( k \)–Means can retrieve these clusters, and (iii) the quality of the resulting clusters is on-par with predictions made by large-scale supervised language identification systems in some cases.

For generating the embeddings, we first strip all punctuation and tokenize by splitting on whitespace. Next, 100-dimensional FastText embeddings are trained on the full corpus yielding the polyglot embeddings. The FastText embeddings use the SkipGram training objective where an input word’s context is predicted. The model is parameterized by a set of real valued vectors (the word-embeddings) for each word in the vocabulary. A full comment (document) embedding can be obtained by normalizing the word-embedding of each of the tokens in the comment and subsequently averaging these word-embeddings.

![Figure 2](http://worldpopulationreview.com/countries/united-kingdom-population)

**Figure 2:** A visualization of the polyglot document-embedding space. The clusters are retrieved with \( k \)-means with \( k \) set to 3 (see, Table 2 for empirical results).

**Qualitative analysis:** A two-dimensional (2D) visualization of the document (comment) embedding space generated through applying

14 [http://worldpopulationreview.com/countries/united-kingdom-population](http://worldpopulationreview.com/countries/united-kingdom-population)
the TSNE algorithm [13] on the computed document embeddings of a random sample of 10,000 comments is shown in Figure 2. We observe three clusters in the visualization. We then run k-Means on these document embeddings setting k to 3 based on this observation. A manual inspection of the clusters reveals that they correspond to (i) Hindi in Roman script (green), (ii) Hindi in Devanagari script (blue), and (iii) English (red).

Quantitative analysis: We next construct a technique for comment language identification. First, each comment’s embedding is obtained by the scheme described above. Next, the value of k for the k-Means algorithm is chosen using a standard heuristic [31] and k-Means is run which yields k clusters. Finally, a sample of 10 comments is drawn from each of the obtained clusters and the dominant language from this sample is assigned to this cluster. In our experience, at least 8 out of 10 comments in the sample were from the dominant language i.e. each of the clusters contains a highly dominant language and the value of k matches the number of languages present in the corpus. A test comment is assigned a language by (i) computing its embedding (as mentioned above), (ii) assigning this comment (embedding) to the cluster whose center is closest, (iii) returning the cluster’s assigned language label.

Table 2: Language written in Roman script is indicated with (E). Percentage of the ground truth assigned this label is indicated for each language. Best metric is highlighted in bold for each language. P: precision, R: recall.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Language</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>0.99</td>
<td>Hindi(E)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>0.99</td>
<td>1.0</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hindi</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>fastTextLangID</td>
<td>0.96</td>
<td>Hindi(E)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>0.99</td>
<td>1.0</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hindi</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>fastTextLangIDRed</td>
<td>0.94</td>
<td>Hindi(E)</td>
<td>1.0</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>0.97</td>
<td>0.71</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hindi</td>
<td>0.97</td>
<td>0.71</td>
<td>0.82</td>
</tr>
<tr>
<td>GoogleLangID</td>
<td>0.95</td>
<td>Hindi(E)</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hindi</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
</tbody>
</table>

We evaluate performance on a held-out set of 200 documents and report precision, recall, F1, and accuracy. 3 languages were discovered by annotators - English, Hindi in Roman script (denoted Hindi(E)), and Hindi written in Devanagari (see, Table 2). Note that, Hindi (mainly spoken in India) and Urdu (mainly spoken in Pakistan) are registers of the same language. Neither our annotators, nor commercial and open source solutions were able to distinguish between the two and thus the Hindi(E) cluster is used to denote both. We compare against two strong supervised baselines - (i) fastTextLangID [17] - a popular open source solution supporting 174 languages, and (ii) GoogleLangID [18] - a commercial solution able to identify close to 100 languages.

Fairness: We first emphasize that the main purpose of comparing against fastTextLangID and GoogleLangID is not to claim our minimally supervised solution is superior to supervised solutions across the board but to demonstrate our solution’s effectiveness in the specific domain of noisy social media texts generated in the Indian subcontinent. A fair performance comparison between the two supervised baselines and our proposed approach is challenging for several reasons. On one hand, due to varying levels of resources, supervised solutions might not be trained to predict the languages (expressed in non-native scripts) in the corpus. The baselines also predict from a larger set of languages. In contrast, our method reveals only those languages observed in the corpus in question - thus a limited set of clusters (labels) is obtained - in most cases this is substantially smaller than the number of languages supported by industrial strength baselines. On the other hand, the baselines are supervised methods that have been trained on vast amounts of annotated data whereas our methods require minimal manual labeling - a critical feature for dealing with corpora featuring low resource languages which are a common occurrence in the Indian subcontinent.

We agree that restricting the baselines to predict only from the smaller set may offset the advantage of our method. The API for fastTextLangID provides an ordered list of all languages that it supports with the confidence score (GoogleLangID does not provide this feature). Let the set of all languages present in a corpus be denoted as L. For a given document, we predict the language belonging to L with the highest confidence score. Suppose the top three predictions for a document from our data set by fastTextLangID are (1) German (predicted with highest confidence), (2) Spanish and (3) Hindi. Since Hindi ∈ L, and German ∉ L, Spanish ∉ L, we consider that the predicted label is Hindi. We denote this new setting as fastTextLangIDfair.

Results are present in Table 2. We observe that our method and the GoogleLangID are able to achieve near-perfect results while both fastTextLangID and fastTextLangIDfair mislabel the Hindi(E) cluster comments as English underscoring the importance of our method. We re-iterate that the purpose of this analysis is to illustrate that when low-resource multilingual settings are encountered, large-scale supervised solutions might not be capable of supporting the desired analyses; our minimal supervision method produces outcomes with reasonable accuracy and avoids imposing significant annotation burdens. Further experiments on additional data sets containing a variety of Indian languages (low-resource languages Bengali and Oriya) and European languages (21 languages) is presented in the Appendix.

Intuition: The SkipGram model used for training the FastText embeddings predicts a word’s context given a word. In a polyglot setting, the likeliest context predicted for a Hindi word is other Hindi words. The embeddings likely reflect this aspect of the language model and thus we see language clusters. We admit that implementation choices like splitting on whitespace (for instance) can preclude some languages so we refrain from making claims about the universality of the technique and present empirical results only on Indian and European languages.

5.2 Temporal trends in pro-peace intent
State of the art sentiment analysis tools typically target domains like movie reviews, product reviews and so on. Prior sentiment analysis research has been performed on political news content [14] and social media responses to humanitarian crises [28]; but to the best of our knowledge, there has been no previous work on war sentiment. Moreover, most of these standard off-the-shelf sentiment analysis tools have been trained on corpora very different from ours. For instance, OpenAI sentiment analysis tool [30] is trained on Amazon e-commerce product reviews. Consequently, off-the-shelf tools are not sufficient in our case. For instance, Stanford CoreNLP (version 3.9.2) sentiment analysis [27], a popular sentiment analysis model, marks the following three examples: [Say no to war.], [War is not a solution.], and [We will nuke you.] as negative, neutral, and positive, respectively. In a conflict-analysis scenario, these three examples should be marked as positive, positive, positive, neutral, positive, negative.
and negative instead. Moreover, we observed that the predicted results are sensitive to punctuation and casing - which cannot be guaranteed in a noisy setting. Hence, we address the challenges in modeling sentiment in our corpus by using a comprehensive manually labeled set of phrases to reveal sentiment. Techniques for analyzing the semantic orientation of text have heavily exploited manually curated lexicons [37, 40, 11, 27]. Following [40, 11], we construct an annotated domain-specific phrase lexicon for mining pro-war and pro-peace intent.

We first identify a set of four high-frequency trigrams expressing collective war/peace intent: [we want peace], [we want war], [we want surgical] (surgical refers to surgical strike), [we want revenge]. These express peace, war, war, and war intents respectively. Out of 9,300,740 unique trigrams, these four trigrams are the 35th, 515th, 875th and 967th in terms of frequency and are the top four collective intent expressing trigrams (trigrams that start with “we” followed by a volitional verb; e.g., [we want], [we need]). A comment that contains $m$ instances of a peace-seeking (war-seeking) phrase receives a positive (negative) score of $m (n)$. The overall score of a comment is $m - n$. The comment expresses peace-seeking intent if the overall score is greater than 0, neutral intent if the overall score is equal to 0 and war-seeking intent if the overall score is less than 0.

We summarize the temporal trends of peace-seeking and war-seeking intent using the four frequently used trigrams in Figure 3(c) and 3(d). We normalize war and peace intent frequencies by the total number of likes or comments received on that day, giving us values in the [0,1] interval, and allowing us to compare activities and sentiment across different days. We measure engagement in terms of comments and likes and plot the overall comment activity (Figure 3(c)) and overall like activity (Figure 3(d)) along with the respective temporal trend plots. As shown in Figure 3(c) and 3(d), the baseline user activity, both in terms of comments and likes, spiked around the IAFPILOT-CAPTURE and IAFPILOT-RELEASE events (nearly 6 times more user engagement on 27th as compared to 15th). Figure 3(c) and Figure 3(d) show that right after PULWAMA, pro-war intent dominated pro-peace intent. Following the pilot’s capture and subsequent release declaration, there was a substantial shift towards pro-peace intent after which, the pro-peace intent generally dominated war-seeking intent. Feb 27th was also the day when Pakistan media reported a meeting between the Pakistan PM and the nuclear warheads body and several news videos discussed the possibility of a nuclear war. Human evaluation on randomly sampled 200 positive comments reveals the following takeaways: in the context of this particular conflict, (1) pro-peace intent spiked when the possibility of a war became real, and (2) the peace-esture by the Pakistan Government possibly influenced this shift as Indians deeply cared about their captured pilot’s safety and appreciated the pilot’s safe return.

In order to widen our coverage, we constructed an extensive lexicon of polarity phrases (sample phrases are listed in Table 3). Overall, we obtained 3,104 annotated phrases as one of: (i) peace-seeking (310 phrases), (ii) war-seeking (278 phrases), or (iii) neutral or unclear (2,516 phrases). Our annotators were instructed to label explicit calls for war and peace. Similar to our previous setting, for a given comment, presence of a peace-seeking phrase contributes +1 to the comment’s score, a war-seeking phrase contributes -1 to the overall score, an neutral phrase contributes a score of 0. The longest matching phrase is considered for computing the sentiment score and all subsumed phrases are ignored. For instance, consider a comment [we want peace but India is not worthy it]; if [we want peace] has score +1 and [we want peace but India is not worthy it] has score 0, [we want peace] is disregarded and the overall contribution from these 2 phrases is 0.

<table>
<thead>
<tr>
<th>Pro-peace</th>
<th>Pro-war</th>
</tr>
</thead>
<tbody>
<tr>
<td>war is not a solution</td>
<td>we are ready for war</td>
</tr>
<tr>
<td>war is not the solution</td>
<td>war is the only solution</td>
</tr>
<tr>
<td>we want peace not war</td>
<td>we are ready to fight</td>
</tr>
<tr>
<td>we dont want war we</td>
<td>we are ready to die</td>
</tr>
<tr>
<td>say no to war</td>
<td>make pakistan</td>
</tr>
<tr>
<td>peace between india and pakistan</td>
<td>make the shit out of</td>
</tr>
<tr>
<td>peace between pakistan and india</td>
<td>want to go to war</td>
</tr>
<tr>
<td>want peace in both countries</td>
<td>war is the only option</td>
</tr>
<tr>
<td>pakistan wants peace with india</td>
<td>wipe out pakistan from the</td>
</tr>
<tr>
<td>war is not a joke</td>
<td>wipe india off the map</td>
</tr>
</tbody>
</table>

Table 3: A random sample of 10 pro-peace and 10 pro-war phrases.

As shown in Figure 3(c) and Figure 3(d) the qualitative trends found in our previous analysis hold. Right after PULWAMA, pro-war intent predominated pro-peace intent and a visible shift was observed on and after Feb 27th. Additionally, our coverage (fraction of comments containing at least one intent-expressing phrase) improved; overall, we obtained 7.25% coverage of comments (20x more than before) and 10.42% (24x more than before) coverage of likes.

**Did many people change their minds?** Unlike YouTube comments, YouTube likes are anonymous and cannot be attributed to individual users. Hence, we focus on the following research question: where there many users who initially clamored for war but later changed their minds? Or, when war became an imminent possibility, did a different sub-population voice their concerns? Analysis reveals the latter case to be true. On our comprehensive intent-expressing phrase set, we found that 4,407 users posted one or more peace-seeking comments, while 7,402 users posted one or more war-seeking comments. 280 users posted both types of comments. The Jaccard index18 between the two user sets was 0.02 indicating low overlap.

**Focused analysis around the peace-spike:** We now focus on a comparative analysis between the two time intervals when war (or revenge) and peace intents were at their respective maximums: a three day period starting on PULWAMA (denoted as war-spike), and a three day period starting on IAFPILLOT-CAPTURE (denoted as peace-spike). We compute the respective unigram distributions $P_{war-spike}$ and $P_{peace-spike}$. Next, for each token $t$, we compute the scores $P_{war-spike}(t) - P_{peace-spike}(t)$, and $P_{war-spike}(t) - P_{war-spike}(t)$ and obtain the top tokens ranked by these scores (indicating increased usage in the respective periods of interest). As listed in Table 4, both war and peace were heavily used tokens during the peace-spike. However, war was predominantly used in the context of peace (e.g., [war is not a solution], [we don’t want war]). Several users also identified themselves as Indian or Pakistani and expressed love for the neighbor country. During the war-spike, demands for revenge, or a surgical strike, or an attack on Pakistan dominated. Heavy use of Kashmir specific keywords during the war-spike and greater emphasis at the country level at the later stage was also consistent with the sequence of events that started as a regional terror attack and snowballed into an international crisis between two nuclear adversaries. We conducted a similar analysis on the set of Hindi comments and our observations align with English corpus.

---

18 defined as $|A \cap B| / |A \cup B|$ for sets $A$ and $B$.
More presence during war-spike
Islam, Kashmiris, Muslim, need, people, religion, Muslims, sad, Modi, China, strike, kill, soldiers, Kashmir, np, revenge, time, surgical, attack, Kashmir

More presence during peace-spike
Pakistan, pilot, war, media, India, pak, peace, Imran, Indian, Pakistani, love, fake, khan, shot, Abhinandan, air, f16, sir, video, mig

Table 4: Biggest shift in token usage in the three day period starting from war-spike and peace-spike.

<table>
<thead>
<tr>
<th>Position</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>eUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>nr phrase</td>
<td>0.30 ± 0.34%</td>
<td>74.44 ± 7.53%</td>
<td>73.50 ± 7.15%</td>
<td>71.30 ± 2.75%</td>
</tr>
<tr>
<td>nr phrase + 1</td>
<td>82.91 ± 2.58%</td>
<td>73.64 ± 4.63%</td>
<td>73.50 ± 2.94%</td>
<td>71.30 ± 2.75%</td>
</tr>
<tr>
<td>nr phrase + 2</td>
<td>82.91 ± 2.58%</td>
<td>75.36 ± 3.90%</td>
<td>74.35 ± 2.24%</td>
<td>71.30 ± 2.75%</td>
</tr>
</tbody>
</table>

Table 5: Hope-speech classifier performance.

5.3 Hope-speech detection

Analyzing and detecting hate-speech and hostility in social media [7, 8, 17] have received considerable attention from the research community. Hate-speech detection and subsequent intervention (in the form of moderation or flagging a user) are crucial in maintaining a convivial web environment. However, in our case where the civilians of two conflicting nations are engaging in heated discussions in a politically tense situation, detecting comments that can potentially diffuse hostility and bring the two countries together has particular importance, for instance by highlighting such comments or otherwise giving them more prominence.

Definition 1: A comment is marked as hope-speech, if it exhibits any of the following:
1. The comment explicitly mentions that the author comes from a neutral country (e.g., [great job thanks from Bangladesh make love not war]), and exhibits a positive sentiment towards both countries in the conflict.
2. The comment explicitly mentions that the author comes from one of the conflicting countries, and exhibits a positive sentiment to an entity (all people, media, army, government, specific professionals) of the other country (e.g., [I am from Pakistan I love India and Indian people]).
3. The comment explicitly urges fellow citizens to de-escalate, to stay calm.
4. The comment explicitly mentions that the author comes from one of the conflicting countries, and criticizes some aspect of the author’s own country (e.g., [I am from India but Indian media very very bad]).
5. The comment criticizes some aspect of both of the conflicting countries.
6. The comment urges both countries to be peaceful.
7. The comment talks about the humanitarian cost of war and seeks to avoid civilian casualties (e.g., [peace is better than war as the price of war is death of innocent peoples]).
8. The comment expresses unconditional peace-seeking intent (e.g., [we want peace]).

If any of the following criteria are met, the comment is not hope-speech:
1. The comment explicitly mentions that the author comes from a conflicting country and expresses no positive sentiment toward the other conflicting country.
2. The comment explicitly mentions that the author comes from a neutral country but takes a position favoring only one of the conflicting countries (e.g., [I m frm Australia I support Pakistan]).
3. The comment actively seeks violence (e.g., [I want to see Hiroshima and Nagasaki type of attack on Pakistan please please]).
4. The comment uses racially, ethnically or nationally motivated
train/validation/test split. On the training set, we train a logistic re-

dependently, and then allowed to discuss and resolve the disagreed la-

together and spread peace. Nobody wants war. We all know what happened in Afghanistan, in Iraq, in Syria. What wounds inflicted by ISIS on the followers of Islam and other reli-
gion. Any religion never spreads hatred. Politicians are spreading hatred. So they also do not have any religion. We are educated mass. We were a single country. No offence for the mistaken identity of the lady. Let’s forget our religion and work for greater peace. Hindustan zindabad and Pakistan zindabad. Terrorism murdabad.

A comment is filled with love. Both the countries do not want any war, and both of us are seeking peace. In spite of that why is this happening? Why the attacks are taking place? Why is there is a heated situation between both the countries. Unfortunately politics is still successful in maintaining the division among Hindustan-Pakistan (divide & rule).

Thank you Pakistan soldiers for good care of our soldier send him back safely to India.

as a pakistani i agree and i think we must pak peoples condemn this war.

Im admiring u sir wen u wer in cricket. U wer the best captain in Pakistan team till now, after dat no one came like u, same way plz try to stop all this war talk ASAP, we love peace as much as u do, sit and talk and dissolve this.... Love n peace from India.

<table>
<thead>
<tr>
<th>Table 6: Random sample of comments in the wild marked as hope-speech by our classifier.</th>
</tr>
</thead>
</table>

Performance in the wild: We randomly sampled 1000 unlabeled comments from each day and ran our hope-speech classifier. Overall, 111 comments were predicted as positives with 94 verified correct by human evaluation (precision: 84.68%). Recall that, simple random sampling uncovered 2.45% of comments exhibiting hope-speech. Hence, our performance in the wild holds promise for substantially reducing manual moderation effort (example comments are presented in Table 6). We further analyzed the sub-categories of positives found in the wild. As shown in Figure 4 our hope-speech classifier was able to find all different sub-categories of positive comments.

6 CONCLUSION

In the era of ubiquitous internet, public opinion on a rapidly evolving global issue can exhibit similar fast-changing behavior, much of which is visible to a very large fraction of internet users. Consequently, this poses an additional challenge to countries with a history of past conflicts as comments inciting hostility may spiral the public opinion towards a stronger pro-war stance. In this work, we define a novel task of hope-speech detection to identify hostility-diffusing content. Extreme web-moderation during periods of strife and tension has included completely disabling internet access in a locality, and report performance on the test set. The experiment was run 100 times on 100 randomly chosen splits. As shown in Table 5 the results indicate that a hope-speech classifier with good precision and recall can be constructed. We admit that off-the-shelf sentiment analysis tools may perform poorly in our task, and it is not a fair compar-

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

- slurs (e.g., pakistani, randia).
- 5. The comment starts the proverbial whataboutism, i.e., we did b because you did a (e.g., [Pakistan started it by causing Pulwama attack killing 44 Indian soldiers]).

Our list of hostility diffusing criteria is not exhaustive and may not cover the full spectrum of hostility diffusing comments. Conse-

quently, we agree that it is possible to have several other reasonable formulations of hope-speech. Also, in a conflict scenario involving more than two conflicting entities, this particular definition may not hold. However, upon manual inspection of the corpus, we found that the definition covers a wide range of potentially hostility-diffusing comments while capturing several nuances.

Hope-speech comment frequency in the wild: On 2000 randomly sampled comments (500 from each week), our annotators found 49 positives (2.45%), 1946 negatives and 5 indeterminate comments. This indicates that detecting hope-speech is essentially a rare positive mining task which underscores automated detection’s importance.

Training set construction using Active Learning: To ensure gen-
eralizability and performance in the wild, it is critical that the training set contains sufficient examples from both classes and captures a wide variety of data points. To ensure this, we divided the corpus into four weekly sub-corpora and sampled uniformly from each of these acknowledging the strong temporal aspect in our data; for a data set consisting of sufficient number of positives and negatives, we employed a combination of Active Learning strategies \[34\] and constructed a data set of 2,277 positives and 7,716 negatives. All rounds of manual labeling were performed by two annotators proficient in English. The annotators were presented with the definition and a small set of examples. 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 \(\kappa\) coefficient across all rounds was 0.82 indicating strong inter-rater agreement).

Features: We considered the following features:

- n-grams up to size 3 following existing literature on text classification \[20\].
- the previously described sentiment score of a comment obtained using our comprehensive set of intent phrases (denoted as I in Table 5).
- the previously described 100-dimensional polyglot FastText embeddings (denoted as FT in Table 5).

Classifier performance: On our final data set, we used a 80/10/10 train/validation/test split. On the training set, we train a logistic regression classifier with L2 regularization with the discussed features and report performance on the test set. The experiment was run 100 times on 100 randomly chosen splits. As shown in Table 5 the results indicate that a hope-speech classifier with good precision and recall can be constructed. We admit that off-the-shelf sentiment analysis tools may perform poorly in our task, and it is not a fair comparison since they are trained for a different domain. However, for the sake of completeness, we ran the Stanford CoreNLP sentiment analyzer on our data set (precision: 27.65%, recall: 41.45%, F1: 33.17%). Our baselines’ stronger performance indicates that the task of hope-speech detection is different from simple sentiment analysis and hence requires a targeted approach.
Our work in detecting hostility-diffusing content may find applications in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES


[5] Mesut Kaya, Guven Fidan, and Ismail H Toroslu, ‘Sentiment analysis of social media data generated in these scenarios as well. We present a thorough analysis of a novel polyglot embedding based language identification module that can be useful in facilitating research on social media data generated in this part of the globe with presence of several low-resource languages.

REFERENCES
7 APPENDIX

YouTube Channels: Table 7 lists the channels we considered.

<table>
<thead>
<tr>
<th>Language</th>
<th>India</th>
<th>Pakistan</th>
<th>World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi News</td>
<td>Aaj TV</td>
<td>Geo News</td>
<td></td>
</tr>
<tr>
<td>India Today</td>
<td>Geo News</td>
<td>NDTV India</td>
<td></td>
</tr>
<tr>
<td>Aaj TV</td>
<td>Dunia News</td>
<td>Daily news</td>
<td></td>
</tr>
<tr>
<td>NDTV &gt;50m</td>
<td>Suttha TV</td>
<td>Al Jazeera</td>
<td></td>
</tr>
<tr>
<td>ABP News</td>
<td>CNN</td>
<td>AT Arabia</td>
<td></td>
</tr>
<tr>
<td>CNN News 18</td>
<td>Zee News</td>
<td>MNN</td>
<td></td>
</tr>
<tr>
<td>India TV</td>
<td>PTV News</td>
<td>Sky News</td>
<td></td>
</tr>
<tr>
<td>BBC World News</td>
<td>Hum News</td>
<td>CNN</td>
<td></td>
</tr>
<tr>
<td>NDTV India</td>
<td>Geo News</td>
<td>FOX News</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: List of popular news channels we considered.

Language Identification: We focus on two low-resource language settings (Bengali and Oriya) and a large mix of well-formed texts from Europe (21 languages, EuroParl data set [15]).

Data sets: We now describe our additional data sets, two of which are collected from the Indian subcontinent, one is a well-known data set of European languages.

- **D_{ABP}**: The ABP Ananda news channel is a Bengali news organization. We crawled the comments on videos uploaded by their YouTube channel [19] and obtained 219,927 comments. Most of the comments are in Bengali, Hindi, and English. Note that internet users in the Indian subcontinent use the Latin script as well as their native script for writing. The use of the Latin script for writing in Hindi and Bengali is significant in this corpus.

- **D_{OTV}**: OTV is an Oriya news network with a popular YouTube channel [20]. We crawled videos from this network and subsequently crawled comments to obtain 153,435 comments. Most of the comments were posted in Oriya, Hindi, and English. Latin script is heavily used alongside the native script for Oriya and Hindi.

- **D_{Euro}**: The Europarl corpus [15] contains 21 languages with well-written text. The processed version is obtained from [23]. 420,000 documents were reserved for training and 210,000 documents were used for test.

Performance on low resource languages: We considered two additional data sets consisting of a mix of languages of which two are low resource languages (Bengali and Oriya). As shown in Table 8 and 9, our performance on the India-Pakistan data set translates to other languages in the Indian subcontinent. We observed that our added fairness criterion marginally improved the performance of fastTextLangID but our method still substantially outperformed fastTextLangID_{fair}. As we already mentioned, we could not construct a similar GoogleLangID_{fair} due to its API’s limitation. However, based on our current observations on fastTextLangID_{fair}, we conjecture that the performance boost would not be substantial.

Performance on EuroParl: Our model’s performance is on-par with fastTextLangID. We did not evaluate against GoogleLangID due to prohibitive costs and it is reasonable to expect very high accuracy due to the clean nature of the corpus. Our method is near-perfect and on-par with fastTextLangID. Our model’s accuracy is 99.9% versus 99.3% for fastTextLangID.

Figure 5: A visualization of the polyglot document-embedding space for the D_{Euro} data set.

Figure 6: A visualization of the polyglot document-embedding space for the D_{OTV} dataset. See Table 9 for empirical results.

Table 9: Performance evaluation on D_{OTV}. Language written in Latin script is indicated with (E). Percentage of the ground truth assigned this label is indicated for each language. Best metric is highlighted in bold for each language.

![Figure 5](https://www.youtube.com/channel/UCV3fFezGHTgXiao3WNc)

![Figure 6](https://www.youtube.com/channel/UCGC5m2BTa6Q)