Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Special Track on AI for Computational Sustainability and Human Well-being

Harnessing Code Switching to Transcend the Linguistic Barrier

Ashiqur R. KhudaBukhsh[1*], **Shriphani Palakodety**[2*] and **Jaime G. Carbonell**[1†]
[1]School of Computer Science, Carnegie Mellon University
[2]Onai
akhudabu@cs.cmu.edu, spalakod@onai.com, jgc@cs.cmu.edu

Abstract

Code mixing (or code switching) is a common phenomenon observed in social-media content generated by a linguistically diverse user-base. Studies show that in the Indian sub-continent, a substantial fraction of social media posts exhibit code switching. While the difficulties posed by code mixed documents to further downstream analyses are well-understood, lending visibility to code mixed documents under certain scenarios may have utility that has been previously overlooked. For instance, a document written in a mixture of multiple languages can be partially accessible to a wider audience; this could be particularly useful if a considerable fraction of the audience lacks fluency in one of the component languages. In this paper, we provide a systematic approach to sample code mixed documents leveraging a polyglot embedding based method that requires minimal supervision. In the context of the 2019 India-Pakistan conflict triggered by the Pulwama terror attack, we demonstrate an untapped potential of harnessing code mixing for human well-being: starting from an existing hostility diffusing *hope speech* classifier solely trained on English documents, code mixed documents are utilized to perform cross-lingual sampling and retrieve *hope speech* content written in a low-resource but widely used language - Romanized Hindi. Our proposed pipeline requires minimal supervision and holds promise in substantially reducing web moderation efforts. A further exploratory study on a new COVID-19 data set introduced in this paper demonstrates the generalizability of our cross-lingual sampling technique.

1 Introduction

Analyzing geopolitical events through the lens of social media is a highly active research domain. From referendums with far-reaching political consequences (e.g., Brexit [Celli *et al.*, 2016]) to sensitive and highly polarizing issues like mass shootings in the US [Demszky *et al.*, 2019], large scale social media analysis has the potential to offer important insights to political and social scientists. While social media discussions lend a great platform to exchange ideas, share opinions and debate issues, tackling online attacks targeted at certain individuals, or communities forms an important modern-day social media challenge to ensure human well-being.

Typical approach to moderate online hate consists of detecting *hate speech* for subsequent moderation. However, one recent line of work argued in favor of identifying positive content in the context of heated online discussions between nuclear adversaries at the brink of waging full-fledged war [Palakodety *et al.*, 2020a]. In a substantial corpus of 2.04 million YouTube comments on videos relevant to the 2019 India-Pakistan conflict triggered by the Pulwama terror attack, Palakodety *et al.* [2020a] advocate the importance of hostility-diffusing comments and define a new task of detecting hostility-diffusing *hope speech*.

While Palakodety *et al.* [2020a] present an important study of modern conflict between two nuclear adversaries with a long history of acrimonious past and grim projection of consequences should there be a full-blown war (100 million projected deaths as forecast by Toon *et al.* [2019]), typical to several studies conducted in linguistically diverse regions, the focus is largely restricted to the English subset of the comments (921,235 English comments posted by 392,460 users). With a combined language base of more than 500 million speakers of Hindi in India and Pakistan as opposed to nearly 250 million English speakers, and a considerable fraction using Romanized Hindi on the web [Gella *et al.*, 2014; Palakodety *et al.*, 2020a], extending this result to the Romanized Hindi subset of comments has understandable benefits. However, such omissions are common in social analyses due to lack of NLP (Natural Language Processing) resources in the native language. Most NLP pipelines are monolingual - state of the art POS (Part-of-Speech) taggers, parsers, or NER (Named Entity Recognition) taggers are rarely trained to handle multi-lingual documents and largely focus on English. Moreover, apart from typical colloquial style social media content, the presence of code mixing [Gumperz, 1982; Myers-Scotton, 1993] – seamless alteration between multiple languages within the same document boundary (e.g., a tweet,

*Ashiqur R. KhudaBukhsh and Shriphani Palakodety are equal-contribution first authors. Ashiqur R. KhudaBukhsh is the corresponding author.

†Deceased 28 February, 2020.

or a comment on a YouTube video) – makes the task substantially more challenging.

While the challenges posed by code switching to downstream analyses are well-documented, in this paper, we focus on a largely under-explored research question: *How can code switching be harnessed for social good and human well-being?*

Our research question is motivated by a simple intuition that a short text document is likely to express a consistent sentiment; if reliable linguistic separation of such code mixed documents can be achieved, the Hindi portion of the comments can be further harnessed to explore similar comments in the Hindi subset for which we require no further training (the *hope speech* classifier we used in this paper is trained on English comments). Effectively, our method uses the Hindi portions of code mixed comments as a seed set to mine similar content authored in Hindi. Our approach presents a compelling case study on how code switching can be harnessed to perform cross-lingual sampling and detect peace-seeking content written in a low-resource language. A reliable system to identify code mixed *hope speech* documents has additional untapped benefits. Intuitively, a code mixed document written in two dominant languages in a linguistically diverse region is likely to be partially accessible to a wider set of audience.

Contributions. Our contributions are the following.

1. **Human well-being:** We focus on the important task of detecting hostility-diffusing *hope speech* [Palakodety *et al.*, 2020a]. Social media is poised to play an increasingly important role in understanding and analyzing modern conflicts [Zeitzoff, 2017]; online discussions between countries with a long history of conflicts are under-studied yet highly important. Additionally, we demonstrate our method's generalizability through a new task introduced in this paper - detecting comments authored in a low-resource language encouraging compliance with COVID-19 health guidance.
2. **Framework:** Code switching is typically viewed as an impediment to effective corpus analysis; to the best of our knowledge, our work is the first to highlight its untapped potential as a bridge between sub-corpora authored in different languages to perform cross-lingual sampling. While the role of mother tongue as a *conversational lubricant* in a code switched environment has been previously studied in educational settings [Butzkamm, 1998], harnessing code switching to effectively sample content from a sub-corpus written in a different language has, to our knowledge, never been explored before.
3. **Machine Learning:** We leverage recent literature in language identification and Active Sampling to sample documents exhibiting high levels of code mixing and provide an end-to-end pipeline to sample from Romanized Hindi starting with a *hope speech* classifier trained on English documents. Our results indicate that our approach considerably reduces manual effort in acquiring *hope speech* written mostly in Romanized Hindi.

Organization of the paper. The rest of the paper is organized as follows. We present literature relevant to our research in Section 2, our problem definition, pipeline and necessary background in Section 3, a detailed description of our methods and performance on *hope speech* detection in Section 4, and an exploratory study on the COVID-19 crisis [Johns Hopkins, 2020] in Section 5. We finally end with our conclusions and proposed extensions in Section 6.

2 Related Work

Code switching has been a widely studied area in linguistics for nearly half a century [Auer, 2013]. While recent work on analyzing the social aspects of code mixing in online communities is gaining importance [Yoder *et al.*, 2017], typically, code switching is viewed as an impediment to downstream NLP analyses and much of the focus in the community is concentrated in token-level language identification and switch point detection for cleaner linguistic separation [Das and Gambäck, 2014; Rijhwani *et al.*, 2017a; Gella *et al.*, 2014]. To the best of our knowledge, harnessing code switching for social good and human well-being has been largely unexplored. Our work draws inspiration from field-work in classroom settings showing how code switching helps students overcome linguistic barriers and how native tongue is used in a code mixed setting as a *conversational lubricant* [Butzkamm, 1998].

Our work focuses on an important domain of online hostility-diffusion between civilians of nuclear adversaries [Palakodety *et al.*, 2020a]. We use several resources presented in the paper (e.g., data set, language identification method with minor modification). However, in the work by Palakodety *et al.* [2020a], the primary focus was mostly restricted to the English subset of comments, whereas in our work, we focus on leveraging an untapped potential of code switching and propose a pipeline to identify hostility-diffusing *hope speech* from the Hindi sub-corpus, a task previously not addressed.

In Gella *et al.* [2014], the importance of a robust token-level language identification system was explored. The study demonstrated that typical document-level language identification systems are a poor fit for code mixed documents. In the context of Indian social media, Gella *et al.* [2014] also provided statistics on the use of Romanized Hindi, and code mixed text revealing significant use. Language preferences to express opinion were further investigated by Rudra *et al.* [2016] revealing that negative opinion is often presented in Hindi. The utility of code switching in improved success rates of Wikipedia edits was studied in [Yoder *et al.*, 2017]. Several studies have addressed challenges in analyzing code mixed text by using a token-level language-identification step in their NLP pipelines [Nguyen and Dogruoz, 2013; Elfardy and Diab, 2013; Rijhwani *et al.*, 2017b]. Rijhwani *et al.* [2017b] in particular presented an HMM-based unsupervised token-level language-identification method to analyze code-switching statistics on social media.

Recent studies have used sentence embeddings for sampling comments similar to a "query" document [Dimovski *et al.*, 2018; Kumar *et al.*, 2019; Palakodety *et al.*, 2020c]. We utilize the polyglot embeddings themselves as sentence embeddings in our nearest-neighbor sampling method.

3 Problem Definition and Background

3.1 Low-resource Language

Low-resource or under-resourced languages lack computational resources such as large corpora (monolingual or parallel) and annotated resources typically needed for NLP methods (e.g., parsers, Named Entity Recognition taggers etc.) [Cieri *et al.*, 2016]. Romanized Hindi or Bengali are two examples of highly prevalent yet low-resource languages.

3.2 Task: *Hope Speech* Detection

We focus on the prediction task of *hope speech* detection in the context of online discussions relevant to the 2019 India-Pakistan conflict [Palakodety *et al.*, 2020a]. Aimed at diffusing hostility, a *hope speech* classifier is a nuanced classifier to detect content that contains a unifying message focusing on the war's futility, the importance of peace, and the human and economic costs involved, or expresses criticism of either the author's own nation's entities or policies, or the actions or entities of the two involved countries (for precise definition with illustrative examples, see [Palakodety *et al.*, 2020a]).

Data set. Our data set, $\mathcal{D}$, consists of 2.04 million comments posted by 791,289 user on 2,890 YouTube videos relevant to this India-Pakistan conflict. Our main focus is on the English and Romanized Hindi subsets denoted as $\mathcal{D}_{en}$ (921,235 comments) and $\mathcal{D}_{h_e}$ (1,033,908 comments), respectively. For the remainder of this paper, we use Romanized Hindi and Hindi interchangeably.

Annotated data set. The *hope speech* classifier is trained on an annotated data set, $\mathcal{D}^{train}_{hope}$, of 2,277 positive and 7,716 negative English comments and an in-the-wild performance (on data not belonging to the training or test set) of 84.68% precision was reported.

3.3 An Illustrative Example

To motivate our intuition, we first provide an illustrative example of a code switched comment exhibiting *hope speech* along with a loose translation. English, Hindi and neutral tokens (e.g., proper nouns, numerals, or technology terms) are color-coded with blue, red and black respectively (color scheme is consistent throughout the paper).

`I am Indian and I say peace is the only solution ankh k badle ankh mangoge toh sari dunya andhi hojayegi`

I am Indian, and I say peace is the only solution; an eye for an eye makes the whole world blind.

In the above example, both the Hindi and English components exhibit peace-seeking intent. Our main goal in this paper is to harness the Hindi components present in these highly code mixed *hope speech* comments to detect *hope speech* in the Hindi sub-corpus. Associated research questions are the following:

- How can we sample code mixed documents?
- How can we harness the Hindi part of a code mixed document to sample *hope speech* from the Hindi portion?

3.4 A Challenging Data Set

Similar to most data sets of noisy, short social media texts generated in a linguistically diverse region, our data set exhibits a considerable presence of out-of-vocabulary (OOV) words, code mixing, and grammar and spelling disfluencies. In addition to these challenges, given that a vast majority of the content contributors do not speak English as their first language, we noticed varying levels of English proficiency in the corpus with a substantial incidence of phonetic spelling errors (e.g., [`thankyou pakusta for hiumaniti no war aman ssnti kayam kare`] loosely translates to *Thank you Pakistan for humanity; let peace prevail.*); 32% of times, the word `liar` was misspelled as `lier`. Since Romanized Hindi does not have any standard spelling (e.g., the word `aman` meaning peace is spelled in the corpus as amun, amaan and aman), a high level of spelling variations added to the challenges.

How hard is it to sample *hope speech*? On a random sample of 1,000 comments from $\mathcal{D}_{h_e}$, our annotators[1] found 18 positives (i.e., 1.8%). This result aligns with results reported by Palakodety *et al.* [2020a] where only 2.45% randomly sampled English comments were marked as *hope speech*. Additionally, a previous study of a multilingual Hindi-English tweet corpus observed that Hindi was more commonly used to express negative sentiment [Rudra *et al.*, 2016]. The minuscule presence of *hope speech* indicates that detecting such content is essentially a rare positive mining task and automated methods are essential.

3.5 Our Pipeline

Research question: *How to harness code switching to sample hope speech from the Hindi subset $\mathcal{D}_{h_e}$?*

A schematic diagram of our pipeline to sample *hope speech* from the Hindi subset, $\mathcal{D}_{h_e}$, is presented in Figure 1. Our pipeline consists of the following steps.

1. Identify the subset, $\mathcal{D}_{cm}$, from $\mathcal{D}_{h_e} \cup \mathcal{D}_{en}$ with substantial code mixing.
2. Run the *hope speech* classifier (trained on annotated English comments $\mathcal{D}^{train}_{hope}$) on $\mathcal{D}_{cm}$ and construct the subset $\mathcal{D}^{hope}$ containing comments predicted as *hope speech*.
3. Construct $\mathcal{D}^{hope}_{h_e}$ transforming each comment in $\mathcal{D}^{hope}$ discarding any tokens not written in Romanized Hindi.
4. Using $\mathcal{D}^{hope}_{h_e}$ as the seed set, retrieve the nearest neighbors in the comment embedding space from $\mathcal{D}_{h_e}$.
5. Manually inspect the obtained sampled comments to detect *hope speech*.

Steps 1, 2, 3, and 4 require minimal manual supervision. Step 5 is the only step that requires substantial manual effort. Our results indicate that we obtained a nearly 10-fold improvement over our baseline (random sampling yields 1.8% *hope speech*).

[1]For all tasks, two annotators proficient in English, Hindi, and Urdu were used. Across all rounds of labeling, the minimum Fleiss' κ measure was high (0.84) indicating strong inter-rater agreement. After independent labeling, differences were resolved through discussion.

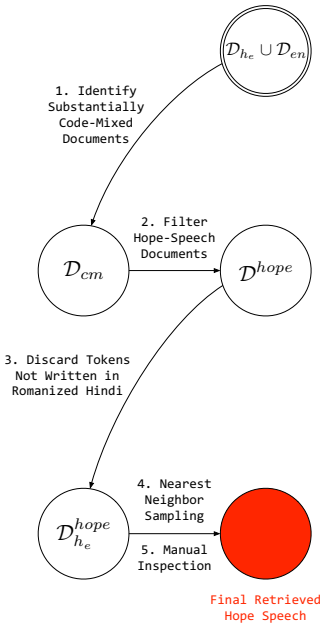

Figure 1: System diagram.

4 Methods and Results

Research question: *how to sample code mixed documents?*

4.1 Code Mixing Index (CMI)

We used a well-known metric to measure the extent of code switching in a document - Code Mixing Index (*CMI*) [Das and Gambäck, 2014]. Essentially, *CMI* measures the presence of a dominant language in a document. Let a document d expressed with k different languages, $\{l_1, \ldots, l_k\}$, and u neutral tokens be represented as a sequence of words: $[w_1, \ldots, w_n]$. Let $\mathcal{L}(w_i)$ return the language of word w_i (or neutral if it is a neutral token). For each language, $\mathcal{N}(l_j)$ denotes the total number of utterances of l_j in the document, i.e., $\mathcal{N}(l_j) = \sum_{i=1}^{i=n} \mathbb{I}(\mathcal{L}(w_i) = l_j)$ where $\mathbb{I}$ is the indicator function. The *CMI* of the document d, *CMI*(d), is measured as:

CMI(d) = $\frac{\sum_{j=1}^{j=k} \mathcal{N}(l_j) - max_i(\mathcal{N}(l_i))}{n-u}$. In the boundary condition, where every word in the document is a neutral token, *CMI* is defined as 0; hence, *CMI*(d) $\in [0, 1)$. A low *CMI* value indicates minimal code switching i.e. the document is almost entirely written in the dominant language. Understandably, when $k = 2$, the highest possible *CMI* is 0.5 indicating equal presence of two component languages. When $\mathcal{L}(.)$ is estimated using a language identification method, we denote the estimated *CMI* of a document as $\widehat{CMI}(d)$.

We now illustrate with an example: [**bilkul sahi baat kahi aapne imran khan saab please please no more war only peace**] (loosely translates to *You've spoken the absolute truth Mr. Imran Khan, please no more war, only peace.*). In this example, $\mathcal{N}(en) = 7, \mathcal{N}(h_e) = 6$, n = 15, and u = 2. Hence, *CMI* of the document is $\frac{7+6-7}{15-2}$ = 0.46. We considered documents with $\widehat{CMI}$ greater than or equal to 0.4 as documents exhibiting significant code mixing.

4.2 Estimating CMI

In order to sample documents with high $\widehat{CMI}$, we need a reliable token-level language identification module. We used the polyglot-embedding based method (denoted by $\hat{\mathcal{L}}_{polyglot}$) proposed by Palakodety *et al.* [2020a]. We chose $\hat{\mathcal{L}}_{polyglot}$ because it requires minimal supervision and is particularly well-suited for noisy social media texts [Palakodety *et al.*, 2020c]. In particular, $\hat{\mathcal{L}}_{polyglot}$ involves obtaining the document embeddings, and then using $k-$means on these embeddings. The method is shown to reveal highly precise monolingual clusters. Previous use-cases of $\hat{\mathcal{L}}_{polyglot}$ were limited to document-level language identification. In our experiments we found that without any significant modification, the technique is capable of token-level language identification with considerable accuracy. Our token-level language identification follows the same method presented by Palakodety *et al.* [Palakodety *et al.*, 2020a]. We consider a token as a single-word document, obtain its embedding, and assign the language of the nearest cluster center in the document embedding space.

Detecting neutral tokens. Neutral tokens are identified using a simple heuristic: for a two-language scenario, a token is marked neutral if it is approximately equidistant from the two respective cluster centers. For a given token, w, let the Euclidean distance of w from the English cluster and Hindi cluster in the comment embedding space be represented as *dist*(w, en) and *dist*(w, h_e), respectively. Let the distance between the two cluster centers be expressed as *dist*(en, h_e). $\hat{\mathcal{L}}_{polyglot}(w)$ = *neutral* iff $w \in (\mathcal{D}_{h_e} \cup \mathcal{D}_{en})$ and $\frac{|dist(w,en) - dist(w,h_e)|}{dist(en,h_e)} \leq \epsilon$.

When ϵ is set to 0.1, our method obtains the following top 20 (ranked by frequency) neutral tokens: `Pakistan`, `he`, `army`, `media`, `Modi`, `Pak`, `Pakistani`, `Kashmir`, `pilot`, `attack`, `video`, `news`, `khan`, `jai`, `2`, `hind`, `Imran`, `Muslim`, `sir`, `1`. These tokens broadly include proper nouns (e.g., `Modi`, `Khan`, `Pakistan`), numerals (e.g., `1`), technical terms (e.g., `video`) and overloaded words (e.g., `he`; `he` translates to *is* in Hindi, and is the third-person singular masculine pronoun in English).

On a data set of 300 comments with gold standard token-level annotation, we found that $\hat{\mathcal{L}}_{polyglot}$ performs token-level language detection with considerable accuracy. As shown in Table 2, the overall accuracy of $\hat{\mathcal{L}}_{polyglot}$ is 88.76%.

Estimating CMI. Once the reliability of token-level language identification by $\hat{\mathcal{L}}_{polyglot}$ is established, we next evaluate the reliability of the estimate for *CMI* (i.e. $\widehat{CMI}$). We first define the subset with substantial estimated code mixing, $\mathcal{D}_{cm}$, as the following: $\mathcal{D}_{cm} = \{d\}$ s.t. $d \in \mathcal{D}_{h_e} \cup \mathcal{D}_{en}$ and $\widehat{CMI}(d) \geq 0.4$, i.e., the document is either part of the

Corpus	*CMI*	$\widehat{CMI}$	Overall RMSE
$\mathcal{D}_{en}$	0.03	0.04	0.05
$\mathcal{D}_{h_e}$	0.10	0.12	
$\mathcal{D}_{cm}$	0.38	0.45	

Table 1: CMI estimation root mean squared error.

		Predicted Label		
		$neutral$	en	h_e
True Label	$neutral$	702	325	144
	en	334	4690	56
	h_e	85	148	3235

Table 2: Confusion matrix of token-level performance evaluation of $\hat{\mathcal{L}}_{polyglot}$ on 300 annotated comments from $\mathcal{D}_{h_e} \cup \mathcal{D}_{en}$.

English or Romanized Hindi subset and its estimated *CMI* is high indicating nearly equal presence of Hindi and English. As shown in Figure 2, $\mathcal{D}_{cm}$ indeed falls in the overlapping region of the English and Hindi cluster. We manually inspected and annotated 1,000 randomly sampled comments from $\mathcal{D}_{cm}$ and found that 95.9% comments exhibited code switching. We further obtained token level consensus labels for 100 randomly sampled comments each from $\mathcal{D}_{cm}$, $\mathcal{D}_{en}$, and $\mathcal{D}_{h_e}$. Table 1 compares the ground truth *CMI* and $\widehat{CMI}$ and demonstrates that we achieved a reasonable approximation of true *CMI* using $\hat{\mathcal{L}}_{polyglot}$.

4.3 Sampling *Hope Speech* From $\mathcal{D}_{h_e}$

Method	Performance
random-Sample($\mathcal{D}_{h_e}$)	1.8%
NN-Sample($\mathcal{D}^{hope}_{h_e}$)	18.59%
NN-Sample($\mathcal{D}^{hope}$)	26.93%
NN-Sample($\mathcal{D}^{hope}_{h_e,+}$)	21.88%
NN-Sample($\mathcal{D}^{hope}_{+}$)	**31.68%**

Table 6: Sampling performance.

Research question: *How to harness the Hindi part of a code mixed document to sample* hope speech *from* $\mathcal{D}_{h_e}$*?*

Once we identify a comment subset with substantial code mixing, $\mathcal{D}_{cm}$, obtaining *hope speech* comments using an off-the-shelf *hope speech* classifier is straight-forward. Out of 36,969 comments in $\mathcal{D}_{cm}$, the classifier predicted a set of 199 comments, $\mathcal{D}^{hope}$, as positives. Upon manual annotation, we

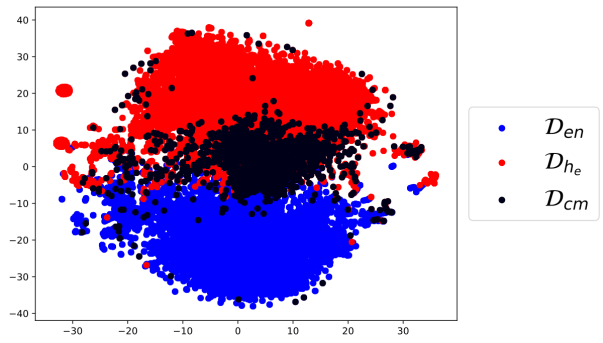

Figure 2: A TSNE [Maaten and Hinton, 2008] plot of the polyglot document-embedding space. The code mixed region (black) lies between the Hindi (red) and English (blue) language clusters.

Algorithm 1: `NN-Sample`$(\mathcal{S}, \mathcal{U})$

Initialization:
$\mathcal{E} \leftarrow \{\}$
Main loop:
foreach comment $c \in \mathcal{S}$ **do**
  $count \leftarrow 0$
  $dist \leftarrow 0$
  while $count \leq size$ **do**
    $neighbor \leftarrow getNearestNeighbor(c, dist)$
    $dist \leftarrow cosineDistance(c, neighbor)$
    if $neighbor \notin \mathcal{E} \cup \mathcal{S}$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{neighbor\}$
      $count \leftarrow count + 1$
    end
  end
end
Output: $\mathcal{E}$

obtained 149 positives (denoted as $\mathcal{D}^{hope}_{+}$), i.e., 74.87% positives. Understandably, due to presence of code switching, the in-the-wild precision in $\mathcal{D}_{cm}$ is lower than previously reported in-the-wild precision of 84.68% in $\mathcal{D}_{en}$ [Palakodety *et al.*, 2020a]. Table 3 lists a subset of randomly sampled comments from $\mathcal{D}^{hope}_{+}$. We noted that the Hindi component of the comments were consistent with the overall sentiment of the comment.

A noisy approximation of the Hindi sub-part of these comments can be obtained by $\hat{\mathcal{L}}_{polyglot}$ through discarding non-Hindi tokens.

I love India I am Pakistani mein amun chahta hon khuda ke waste jang nai peace peace peace

I love India, I am Pakistani. I want peace for God's sake, not war, peace peace peace.

For instance, the above comment is transformed into [**mein amun chahta hon khuda ke jang nai**] (loosely translates to *I want peace for God not war*) when we discard non-Hindi tokens using $\hat{\mathcal{L}}_{polyglot}$. `Waste` is both a valid English and Hindi word (meaning *sake*), and the language detector makes an error in correctly predicting it. We admit that it is possible to use more sophisticated methods to extract Hindi that consider context (e.g., considering context to assign label to a fence word) and possibly squeeze more performance out of it. However, we are primarily interested in establishing a blue-print for harnessing code switching for social good and testing the robustness of our pipeline without resorting to performance-driven engineering. In every step of our pipeline, a better-performing algorithm (e.g., better language detection module, sophisticated method to extract Hindi, more powerful comment embeddings, further effective sampling technique) can be plugged in without disturbing the flow and with a possibility of performance improvement.

Active Sampling. Once we extract the Hindi sub-parts of $\mathcal{D}^{hope}$ (denoted as $\mathcal{D}^{hope}_{h_e}$), our next task is to find comments in $\mathcal{D}_{h_e}$ that are similar to the Hindi sub-part. To this end,

Code switched *hope speech*	Loose translation
I am Pakistani agar ap dono country ny war karne ha to gurbat khatam karnay ke war karo dono countries bht gareeb hain plzz dont do war war is not solution of peace	*I am Pakistani. If both countries have to wage a war, wage a war to end poverty; both countries are very poor. Please, do not war, war is not solution of peace.*
please media walo nafrat phailana chhod do we want peace only jai hind	*Please media folks, stop spreading hate. We want peace; hail India!*
absolutely right i think ab netao kee jung ko ham dono mumalik ne rad karna hai we people of both countries want peace peace and peace	*Absolutely right. I think politicians' war has to prevented by common people of both countries. We people of both countries want peace, peace, and peace.*

Table 3: Random sample of code mixed *hope speech* obtained by *hope speech* classifier run on $\mathcal{D}_{cm}$.

Code switched *hope speech*	Loose translation
bhai ap bhi khuch rahiye allah apki har farmaish puri kare and I repeat again bhai I love you all my dear brothers and sisters mujhe ekh dusre se indian or pakistaani keh kar bulana bilkul pasand nahi hum sab bhai or bhen hai or rahenge we are good humans of earth	*Brother, you also be happy. May Allah grant all your wish and I repeat again, brother, I love you all my dear brothers and sisters. I don't like to identify each other as Indian or Pakistani, we are all brothers and sisters, we are good humans of earth.*
galiyan dene se kya hoga beach me to begunah awam mare gii orr ham jo jang jang krte henn kha jang itnii asan he no this war is end of the world because Ind an Pak is newclear states	*Nothing will come out of abusing, we will cry for war while innocent civilians will die. Who said that war is easy? No, this war is end of the world because Ind and Pak are nuclear states.*
daikhou dosto apaas me baahss bazii maat kro plz such me I have love for both Pakistan and India bus apaas me muhabaatey rakhou I m student of 9th lakn me muhabaat chataa hu donoo countries me choroo fazoul ki nafrateey love you Pakistan and my neighbour country India	*Look friends, stop quarrelling among each other. Please, for real, I have love for both Pakistan and India, just harbor love between each other. I am a student of 9th grade, but I want love between both countries. Leave this useless hate, love you Pakistan and my neighbor country India.*

Table 4: Random sample of *hope speech* obtained through *NN-Sample*($\mathcal{D}^{hope}$).

we use a recently-proposed Active Sampling algorithm which samples nearest neighbors in the comment embedding space to identify rare positives [Palakodety *et al.*, 2020c]. Our choice of this Active Sampling technique is motivated by its effectiveness in mining rare positives and reported robustness to spelling variations which is particularly critical because our corpus contains noisy social media texts and Romanized Hindi does not have standard spelling rules. Following [Palakodety *et al.*, 2020c], we used cosine distance of the embeddings as the distance measure. Our sampling algorithm is described in Algorithm 1. This algorithm takes a seed set, $\mathcal{S}$, and a sample pool $\mathcal{U}$ as inputs and outputs a set, $\mathcal{E} \subset \mathcal{U}$, containing nearest neighbors of $\mathcal{S}$ in the comment-embedding space. Initially, $\mathcal{E}$ is an empty set. At each step, we expand $\mathcal{E}$ with nearest neighbors that are not present in the expanded set or the seed set. The function *getNearestNeighbor*(*c, dist*) returns the comment in $\mathcal{U}$ with minimum distance greater than or equal to *dist*. The *size* parameter is set to 5, i.e., for each comment, we add five unique nearest neighbors. We set $\mathcal{U}$ to $\mathcal{D}_{h_e}$ since we are interested in detecting *hope speech* in Hindi.

Baselines. Recall that, a random sample of 1,000 comments from $\mathcal{D}_{h_e}$ only yielded 1.8% positives which is our primary baseline method (denoted as *random-Sample*($\mathcal{D}_{h_e}$)).

Table 6 compares the performance of our sampling method against the baseline (we do not explicitly mention $\mathcal{U}$ which is consistently set to $\mathcal{D}_{h_e}$ across all *NN-Sample* methods). We obtained substantial improvement over the baseline. Both *NN-Sample*($\mathcal{D}^{hope}_{h_e}$) and *NN-Sample*($\mathcal{D}^{hope}$) require human inspection only at the last step of our pipeline. Our results indicate that our approach can substantially reduce manual effort in acquiring *hope speech*. Effectively, we sampled *hope speech* from a Hindi corpus simply relying on a classifier

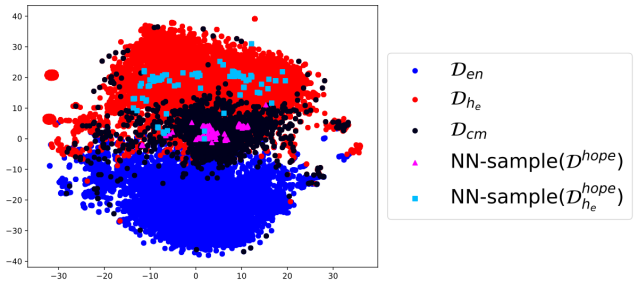

Figure 3: A 2D visualization showing the sampling results against the embedding space. Discarding non-Hindi tokens retrieves documents with low *CMI* written mostly in Romanized Hindi.

trained on English comments and harnessing code switching as a bridge between the Hindi and English sub-corpora. In all steps of the pipeline, we perform noisy approximations in estimating *CMI*, extracting Hindi sub-parts of comments and of course, detecting *hope speech*. If we introduce little more supervision and instead expand the manually annotated *hope speech* set $\mathcal{D}^{hope}_{+}$, as expected, our performance improved. Our results indicate that using minimal manual supervision we can sample with more than 30% accuracy from the Hindi subset $\mathcal{D}_{h_e}$.

Research question: *What is the benefit of extracting the Hindi sub-part?* Both *NN-Sample*($\mathcal{D}^{hope}_{h_e}$) and *NN-Sample*($\mathcal{D}^{hope}_{h_e,+}$) are outperformed by corresponding sampling methods *NN-Sample*($\mathcal{D}^{hope}$) and *NN-Sample*($\mathcal{D}^{hope}_{+}$), respectively (see, Table 6). We were curious to analyze if extracting the Hindi allows sampling from the sub-region of $\mathcal{D}_{h_e}$ mostly written in pure Hindi. As shown in Table 7, without remov-

Sampled *hope speech*	Loose translation
jung kisi maslay ka hal nie aman qaim kro	*War won't solve any problems, restore peace.*
beshak India aur Pakistan ko bhaith kar baat ko suljana chahiye kyunke is ladaye me humare desh ke fouji jo bevajaah shahid ho rahe h aur Pakistan ke fouji jo bevajaah shahid ho rahe h is me na hi mantriyo ka koi nuksaan h na hi kisika ...	*Of course, India and Pakistan should sit together and solve this through dialogue. In this war, ministers and others stand to lose nothing from the pointless deaths of Indian and Pakistani soldiers...*
mein ek rajput hun aur hum kbhi nh chahty k donu mulk apse mein lary phely hum ek thai phir juda huwa kuch intah pasand log nhi chaty k khoon khraba na hon	*I am a Rajput, I never want fight between the two countries; We were one country before the partition, only a handful of extremists want bloodshed.*

Table 5: Random sample of *hope speech* obtained through *NN-Sample*($\mathcal{D}_{h_e}^{hope}$).

Method	$\widehat{CMI}$
$\mathcal{D}_{h_e}$	0.12
$\mathcal{D}_{en}$	0.04
$\mathcal{D}_{hope}^{train}$	0.03
$\mathcal{D}_{cm}$	0.44
NN-Sample($\mathcal{D}_{hope}^{h_e}$)	0.05
NN-Sample($\mathcal{D}_{hope}$)	0.43

Table 7: $\widehat{CMI}$ comparison.

ing the non-Hindi part, *NN-Sample*(.) (intuitively) yielded a set with high level of code mixing. Nearest neighbors of a code switched comment are likely other code switched comments. However, sampling using just the Hindi sub-part yielded substantially less code mixing - Table 5 shows considerably less code-mixing than Table 4. Our intuition that removing non-Hindi tokens is crucial for sampling from the low *CMI* region of $\mathcal{D}_{h_e}$ is further supported by Figure 3 wherein *NN-Sample*($\mathcal{D}_{h_e}^{hope}$) has a better spread over $\mathcal{D}_{h_e}$ while *NN-Sample*($\mathcal{D}^{hope}$) is mostly located in the code mixed region.

5 COVID-19 Health Guidelines Compliance

Research question: *Can our method generalize to other domains?* In this section, we introduce a new data set relevant to the novel COVID-19 pandemic [Johns Hopkins, 2020] and present an exploratory study on a new task - detecting comments encouraging compliance with COVID-19 health guidelines. Using minor modifications to our proposed pipeline, starting with a handful of English example comments, we show that it is possible to perform cross-lingual sampling and detect similar content in Romanized Hindi.

Data set. Our data set consists of 3,144,988 comments on 44,888 YouTube videos uploaded by 14 highly-subscribed Indian news outlets (previously used in [Palakodety *et al.*, 2020b]) between 30 January, 2020[2] and 10 April, 2020. Using $\hat{\mathcal{L}}_{polyglot}$, we obtained 771,035 English comments (denoted by $\mathcal{D}_{en}^{covid}$) in and 1,720,703 comments in Romanized Hindi (denoted by $\mathcal{D}_{h_e}^{covid}$).

Task. Our goal is to find comments in $\mathcal{D}_{h_e}^{covid}$ exhibiting compliance to health guidelines. Health guidelines were regularly revised during this period; we narrowed our focus on the following five guidelines recommended by CDC[3] (1) maintaining social distancing (2) avoiding public gatherings (3) staying home when sick (4) covering coughs and sneezes and (5) washing hands regularly. An example code-switched comment is presented below.

ap ka ghar se nikla ek kadam desh ke karodo logo ki kurbani pe pani dal dega so be alert and aware about our duty as citizens of India we should take oath to win against this corona and bad time

A single step out of your house will nullify the sacrifice of millions of citizens. So be alert and aware about our duty as citizens of India; we should take oath to win against this Corona and bad time.

Our work is related to Mate *et al.* [2020] in its shared focus of COVID-19 analysis in India; however Mate *et al.* [2020] investigated policy design questions, whereas we are primarily interested in mining relevant content in a low-resource language.

Unlike the previously discussed task of *hope speech* detection, we do not have access to a content classifier that can detect comments encouraging compliance with COVID-19 health guidelines. We instead start with a handful of example English comments specified by our annotators and aim to retrieve similar comments authored in Romanized Hindi. Our comments subset, $\mathcal{A}_{target}$, consists of the following five comments: `[Please maintain social distancing]`, `[Please avoid public gatherings]`, `[Please stay at home when sick]`, `[Please cover your coughs and sneezes]`, `[Please wash your hands regularly]`. Starting with $\mathcal{A}_{target}$, we aim to retrieve similar comments in $\mathcal{D}_{h_e}^{covid}$. Note that, in this new setting, the English phrases are not sourced from the corpus but are authored by annotators. In contrast, the previous task of *hope speech* detection utilized a classifier to obtain the initial set of comments from $\mathcal{D}_{cm}$.

Recall that, the pipeline for *hope speech* discovery in Romanized Hindi starts with a set of code mixed documents discovered by a classifier. The English tokens are discarded to formulate Romanized Hindi phrases and then the Romanized Hindi sub-corpus is queried for similar samples using these phrases. The end result is *hope speech* documents in Romanized Hindi. In this task, we start with annotator authored English documents encouraging compliance with guidelines. We first sample documents in the corpus semantically similar to these authored documents using *NN-Sample*($\mathcal{A}_{target}$, $\mathcal{D}_{en}^{covid}$). This yields documents that encourage compliance au-

[2]First COVID-19 positive case was reported in India on this day.

[3]https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

Sampled comments encouraging compliance with health guidelines	Loose translation
sab log party karna band karo na kuchh din ke liye party ni karoge to ni ji paoge ka	*Stop partying for a few days, will you die if you don't party?*
...sirf log jagruta se hi kuch hadh tak bach sak ta hai jaise ki haath senetaiz se dhole aur musk peheny aur saaf sutra rahe...	*Only public awareness can save us somewhat, for instance washing hands with a sanitizer, wearing a mask, maintaining hygiene...*
shab e barat main ibabat apne apne gharon main hi karen ... shatan bimari ke khilaf insan ki larai ka saath den ajmer sharif	*Please offer your prayers on Shab-e-baarat at home ...Ajmer Sharif, please help in this fight against this evil disease.*

Table 8: Random sample of comments obtained through our method.

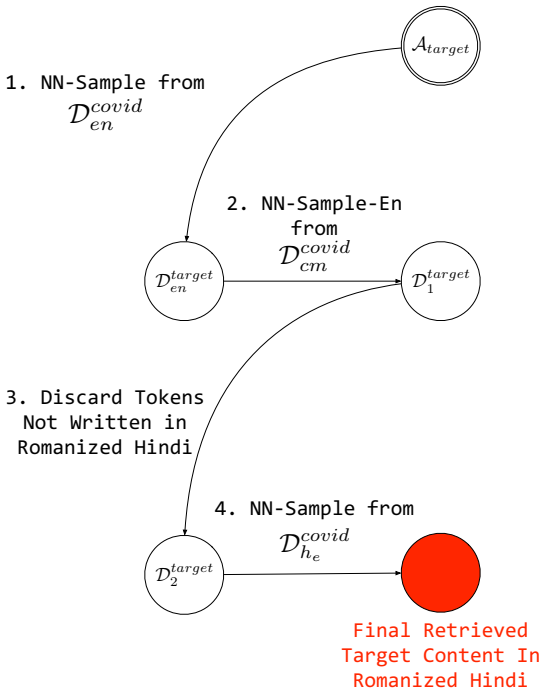

Figure 4: Modified pipeline to tackle absence of a content classifier.

thored in English and are *present in the corpus*. A next round of sampling is conducted to yield code mixed documents similar to these English documents (*NN-Sample-En*(.)). *NN-Sample-En*(.) is identical to *NN-Sample*(.) presented in Algorithm 1 with only one modification: it computes semantic similarity between an English document and a code mixed document by discarding the non-English tokens from the latter. At this stage, we have code mixed documents sourced from the corpus and the rest of the pipeline is similar to the previous task - any tokens not written in English are deleted, the Romanized Hindi portions are retained and used for sampling from the Romanized Hindi sub-corpus yielding comments authored in Romanized Hindi that encourage compliance with health guidelines. The added sampling phases in this particular task allow us to compensate for the lack of a classifier. We first use the annotator authored documents to obtain documents in the corpus that exhibit our desired properties and then utilize the pipeline developed earlier in this paper. Figure 4 shows the system diagram for this task.

For all rounds of nearest neighbor sampling, $size$ was set to 5 yielding 625 comments sampled from $\mathcal{D}_{h_e}^{covid}$. Upon annotation, we found 14.88% positives. A random sample of 625 comments from $\mathcal{D}_{h_e}^{covid}$ yielded 2.88% positives. Hence, even under this additional resource constraint, our pipeline obtained more than 5-fold performance improvement over a random baseline. Table 8 lists a sample of our retrieved comments that shows our pipeline obtained content authored mostly in Romanized Hindi (average $\widehat{CMI}$ of 0.05).

6 Conclusions and Future Work

In NLP literature, typically, code switching is viewed as an impediment to downstream analyses. In this paper[4], we first raise a novel proposition that code switching can be harnessed for social good and human well-being. We utilise it as a bridge between a resource-rich and a low-resource language to reduce annotation efforts in the latter while leveraging resources tailored to the former. Our approach is appealing for its minimal supervision requirements. In the context of hostility diffusing *hope speech* comments, our methods can be used to broaden the reach of such content overcoming the varied language skills of linguistically diverse regions and transcending language barriers. In relation to the novel COVID-19 pandemic, we utilize a small set of annotator authored English phrases encouraging compliance with health guidelines and retrieve similar Hindi content. Our method holds significant promise in addressing resource gaps across widely used languages. Future lines of research include (1) exploring a broader range of language pairs (2) investigating applicability in additional domains and (3) evaluating performance improvement through pipeline modifications.

Acknowledgements

We thank Rupak Sarkar for his assistance in compiling our COVID-19 data set.

This paper is dedicated to Professor Jaime G. Carbonell and his scientific contributions to the fields of Machine Learning, Natural Language Processing, and Artificial Intelligence.

References

[Auer, 2013] Peter Auer. *Code-switching in conversation: Language, interaction and identity*. Routledge, 2013.

[Butzkamm, 1998] Wolfgang Butzkamm. Code-switching in a bilingual history lesson: The mother tongue as a conversational lubricant. *International Journal of Bilingual Education and Bilingualism*, 1(2):81–99, 1998.

[4]Resources and additional details are available at: https://www.cs.cmu.edu/~akhudabu/CodeSwitching2020.html

[Celli *et al.*, 2016] Fabio Celli, Evgeny Stepanov, Massimo Poesio, and Giuseppe Riccardi. Predicting brexit: Classifying agreement is better than sentiment and pollsters. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 110–118, 2016.

[Cieri *et al.*, 2016] Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, 2016.

[Das and Gambäck, 2014] Amitava Das and Björn Gambäck. Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, 2014.

[Demszky *et al.*, 2019] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of NAACL-HLT 2019*, pages 2970–3005. ACL, June 2019.

[Dimovski *et al.*, 2018] Mladen Dimovski, Claudiu Musat, Vladimir Ilievski, Andreea Hossmann, and Michael Baeriswyl. Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4019–4025. AAAI Press, 2018.

[Elfardy and Diab, 2013] Heba Elfardy and Mona T. Diab. Sentence level dialect identification in arabic. In *ACL*, 2013.

[Gella *et al.*, 2014] Spandana Gella, Kalika Bali, and Monojit Choudhury. "ye word kis lang ka hai bhai?" testing the limits of word level language identification. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 368–377. NLP Association of India, December 2014.

[Gumperz, 1982] John J Gumperz. *Discourse strategies*, volume 1. Cambridge University Press, 1982.

[Johns Hopkins, 2020] CSSE Johns Hopkins. Coronavirus covid-19 global cases, 2020.

[Kumar *et al.*, 2019] Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of NAACL-HLT 2019*, pages 3609–3619, 2019.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Mate *et al.*, 2020] Aditya Mate, Jackson A Killian, Bryan Wilder, Marie Charpignon, Ananya Awasthi, Milind Tambe, and Maimuna S Majumder. Evaluating covid-19 lockdown policies for india: A preliminary modeling assessment for individual states. *Available at SSRN 3575207*, 2020.

[Myers-Scotton, 1993] Carol Myers-Scotton. Dueling languages: Grammatical structure in code-switching. claredon, 1993.

[Nguyen and Dogruoz, 2013] Dong-Phuong Nguyen and A. Seza Dogruoz. Word level language identification in online multilingual communication. In *Proceedings of EMNLP 2013*, pages 857–862. ACL, 2013.

[Palakodety *et al.*, 2020a] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Hope speech detection: A computational analysis of the voice of peace. In *Proceedings of ECAI 2020*, page To appear, 2020.

[Palakodety *et al.*, 2020b] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Mining insights from large-scale corpora using fine-tuned language models. In *Proceedings of the Twenty-Fourth European Conference on Artificial Intelligence (ECAI-2020)*, page To appear, 2020.

[Palakodety *et al.*, 2020c] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Voice for the voiceless: Active sampling for finding comments supporting the rohingyas. In *Proceedings of AAAI 2020*, page To appear, 2020.

[Rijhwani *et al.*, 2017a] Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of ACL 2017*, pages 1971–1982, 2017.

[Rijhwani *et al.*, 2017b] Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the ACL 2017*, pages 1971–1982. ACL, July 2017.

[Rudra *et al.*, 2016] Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of EMNLP 2016*, pages 1131–1141, 2016.

[Toon *et al.*, 2019] Owen B Toon, Charles G Bardeen, Alan Robock, Lili Xia, Hans Kristensen, Matthew McKinzie, RJ Peterson, Cheryl S Harrison, Nicole S Lovenduski, and Richard P Turco. Rapidly expanding nuclear arsenals in pakistan and india portend regional and global catastrophe. *Science Advances*, 5(10):eaay5478, 2019.

[Yoder *et al.*, 2017] Michael Yoder, Shruti Rijhwani, Carolyn Rosé, and Lori Levin. Code-switching as a social act: The case of Arabic Wikipedia talk pages. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 73–82. ACL, August 2017.

[Zeitzoff, 2017] Thomas Zeitzoff. How social media is changing conflict. *Journal of Conflict Resolution*, 61(9):1970–1991, 2017.