On NLP Methods Robust to Noisy Indian Social Media Data

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

Much of the computational social science research focusing on issues faced in developing nations concentrates on web content written in a world language often ignoring a significant chunk of a corpus written in a poorly resourced yet highly prevalent first language of the region in concern. Such omissions are common and convenient due to the sheer mismatch between linguistic resources offered in a world language and its low-resource counterpart. However, the path to analyze English content generated in linguistically diverse regions, such as the Indian subcontinent, is not straight-forward either. Social science/AI for social good research focusing on Indian sub-continental issues faces two major Natural Language Processing (NLP) challenges: (1) how to extract a (reasonably clean) monolingual English corpus? (2) How to extend resources and analyses to its low-resource counterpart? In this paper[1], we share NLP methods, lessons learnt from our multiple projects, and outline future focus areas that could be useful in tackling these two challenges. The discussed results are critical to two important domains: (1) detecting peace-seeking, hostility-diffusing *hope speech* in the context of the 2019 India-Pakistan conflict (2) detecting user generated web-content encouraging COVID-19 health compliance.

1 Introduction

Analyzing different aspects of the society through the lens of social media is a highly active research domain [Koutra *et al.*, 2015; Celli *et al.*, 2016; Demszky *et al.*, 2019]. From referendums [Celli *et al.*, 2016] to modern conflicts [Palakodety *et al.*, 2020a], web-scale social media analyses allow us to aggregate opinions of thousands like never before. However, much of the computational social science research focusing on developing nations such as India typically concentrates on web content written in a world language [Kagan *et al.*, 2015; Roy *et al.*, 2017; Jaidka *et al.*, 2019; Palakodety *et al.*, 2020a; Palakodety *et al.*, 2020b]. Such omissions are common and convenient due to the stark contrast between linguistic resources available in a world language such as English and resources available for a highly prevalent yet poorly resourced Romanized Hindi (more than 450 million speakers in India and Pakistan).

The path to analyze English content generated in the linguistically diverse Indian subcontinent is not straight-forward either. Gathering data from YouTube [HindustanTimes, 2019], by far the most popular Indian social media platform, requires a reliable language identifier that can separate (reasonably clean) monolingual corpora. In this paper, we focus on two important research questions that we believe are crucial for effective analysis of Indian social media: (1) how to extract a (reasonably clean) monolingual English corpus? (2) How to extend resources and analyses to its low-resource counterpart? We share NLP methods and lessons learnt through summarizing our work on two important domains: (1) detecting hostility-diffusing, peace-seeking *hope speech* in the context of the 2019 India-Pakistan conflict [Palakodety *et al.*, 2020a] and (2) detecting user generated web-content encouraging COVID-19 [Johns Hopkins, 2020] health compliance [KhudaBukhsh *et al.*, 2020].

2 Background and Data

2.1 Tasks

We focus on the following two tasks:

- ***Hope speech* detection**: Introduced in [Palakodety *et al.*, 2020a] in the context of heated political discussions between two nuclear adversaries at the brink of a full-fledged war, our work advocates the importance of detecting hostility-diffusing, peace-seeking *hope speech*. India and Pakistan have a long history of conflicts with four major wars and many skirmishes; a recent study reported a grim forecast of 100 million deaths should there be a full-fledged war be-

[*] Ashiqur R. KhudaBukhsh and Shriphani Palakodety are equal-contribution first authors.

[†] Deceased 28 February, 2020.

[1] Accepted at **ECAI 2020**, the 24th European Conference on Artificial Intelligence. Camera ready link: http://ecai2020.eu/papers/869_paper.pdf.
Accepted at **IJCAI-PRICAI 2020**, the International Joint Conference on Artificial Intelligence-Pacific Rim International Conference on Artificial Intelligence. Camera ready link: https://www.ijcai.org/Proceedings/2020/0602.pdf.

tween these two nuclear powers [Toon *et al.*, 2019]. 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 a precise definition, see [Palakodety *et al.*, 2020a]). We see this work as a part of the recent trend of *counter speech* research [Benesch *et al.*, 2016; Benesch, 2014; Mathew *et al.*, 2018; Palakodety *et al.*, 2020c].

- **COVID-19 health guidelines compliance detection**: Presented in [KhudaBukhsh *et al.*, 2020], this task involves detecting comments in exhibiting compliance to health guidelines. We focus on the following five CDC-recommended guidelines[2] (1) maintaining social distancing (2) avoiding public gatherings (3) staying home when sick (4) covering coughs and sneezes and (5) washing hands regularly.

2.2 Data

We consider the following two data sets:

- $\mathcal{D}_{hope}$ consists of 2.04 million comments posted by 791,289 user on 2,890 YouTube videos relevant to the 2019 India-Pakistan conflict [Palakodety *et al.*, 2020a].
- $\mathcal{D}_{covid}$ consists of 3.14 million comments on 44,888 YouTube videos uploaded by 14 highly-subscribed Indian news outlets between 30 January, 2020[3] and 10 April, 2020 [KhudaBukhsh *et al.*, 2020].

3 Methods and Results

3.1 Language Identification

Research question: *How to extract a (reasonably clean) monolingual corpus from a multilingual data set?*

While language identification of well-formed text is a nearly-solved problem, the difficulty in identifying language in a noisy social media setting is a non-trivial challenge [Bergsma *et al.*, 2012; Gella *et al.*, 2014; Lui and Baldwin, 2014; Jaech *et al.*, 2016; Jauhiainen *et al.*, 2019]. Nearly 90% of the social media content in Indic regional languages uses Roman script instead of their respective traditional scripts [Gella *et al.*, 2014]. Indic languages when expressed in Roman script, do not have any consensus spelling rule, which further exacerbates the challenge. However, separating out portions of the corpus written in distinct languages is a critical step for most downstream analyses. We now describe a simple yet effective and highly annotation efficient method, dubbed $\widehat{\mathcal{L}}_{polyglot}$, to address this language identification task proposed in [Palakodety *et al.*, 2020a].

$\widehat{\mathcal{L}}_{polyglot}$: Polyglot word embeddings are obtained when a single Skip-gram model [Mikolov *et al.*, 2013; Bojanowski *et al.*, 2017] is trained on a multilingual corpus. Polyglot word embeddings have been found to be useful in several downstream tasks [Mulcaire *et al.*, 2018; Mulcaire *et al.*, 2019a; Mulcaire *et al.*, 2019b]. In our work [Palakodety *et al.*, 2020a], we propose a polyglot embedding-based language identification algorithm that requires minimal supervision. Our algorithm leverages a novel observation (also reported first in our work) that when polyglot document embeddings are clustered using a standard k-means algorithm, precise monolingual clusters are obtained. Assigning language labels to such clusters is a trivial task and requires labeling only a handful of comments randomly sampled from the cluster. Once the language clusters are labeled, for any new document, the language detection task simply boils down to obtaining the document embedding and then assigning the document a language label same as its nearest cluster center.

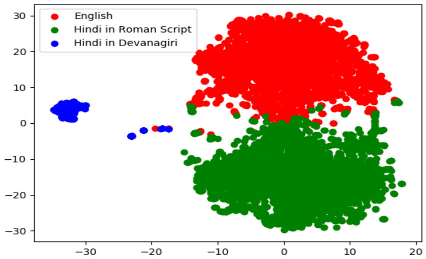

Figure 1: A t-SNE [Maaten and Hinton, 2008] 2D visualization of the polyglot document-embedding space. The clusters are retrieved with k-means with k set to 3. 30 annotated comments (10 per cluster) were required to assign language labels.

Annotation efficiency: As shown in Table 1, $\widehat{\mathcal{L}}_{polyglot}$ marginally outperforms `GoogleLangID`, a really strong baseline. As much as performance is a key objective, in a low-resource setting, annotation efficiency is also equally important [Joshi *et al.*, 2019]. $\widehat{\mathcal{L}}_{polyglot}$ is highly annotation efficient. Using $\widehat{\mathcal{L}}_{polyglot}$, we obtained high-quality language labels for the entire 2.04 million comments in $\mathcal{D}_{hope}$ with 30 annotated comments. When we compare against commercial solutions, we do not claim that $\widehat{\mathcal{L}}_{polyglot}$ is better than the baselines across the board. Rather, we seek to attract the attention of the research community to the surprisingly minimal annotation requirement of $\widehat{\mathcal{L}}_{polyglot}$ and what that entails. Existing systems have minimal support for Romanized Indic languages which in fact, are the most prevalent forms of web-expression in Indian social media [Gella *et al.*, 2014]. For example, `FastTextLangID` provides no support for any Romanized Indic language, and `GoogleLangID` does not support Romanized Odia and Assamese. Assam has been a center for political debates and unrest in recent times [BBC, 2020]. Social scientists interested in doing research on the current crisis can greatly benefit from our language identification method.

Adoption: $\widehat{\mathcal{L}}_{polyglot}$ has been successfully used in our other projects such as: (1) a *counter speech* analysis on the Rohingya refugee crisis [Palakodety *et al.*, 2020c] and (2) an analysis of the 2019 Indian General Election [Palakodety *et al.*, 2020b] through the lens of `BERT` [Devlin *et al.*, 2019].

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

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

Method	Accuracy	Language	P	R	F1
$\widehat{\mathcal{L}}_{polyglot}$ [Palakodety *et al.*, 2020a]	**0.99**	Hindi (E) (52.5%)	**1.0**	**0.98**	**0.99**
		English (46.5%)	**0.99**	**1.0**	**0.99**
		Hindi (1%)	**1.0**	**1.0**	**1.0**
`fastTextLangID` [Facebook, 2016]	0.48	Hindi (E) (52.5%)	**1.0**	0.01	0.02
		English (46.5%)	0.55	**1.0**	0.71
		Hindi (1%)	**1.0**	**1.0**	**1.0**
`GoogleLangID` [Google, 2020]	0.96	Hindi (E) (52.5%)	0.97	0.94	0.96
		English (46.5%)	0.97	0.97	0.97
		Hindi (1%)	0.4	**1.0**	0.57

Table 1: 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.

3.2 Resource Transfer to Romanized Hindi

Research question: *Given resources (e.g., a content classifier or labeled examples) present in English, how can we extend resources and analyses to a low-resource Romanized Hindi?*

Code switching (or code mixing) – the seamless alteration between multiple languages within the same document boundary – has been widely studied in linguistics [Auer, 2013]. Typically, code switching is viewed as an impediment to downstream NLP analyses, but our approach [KhudaBukhsh *et al.*, 2020] views it as an asset that enables transfer from high-resource to lower resource languages. Our cross-lingual sampling technique is guided 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 portion of the content authored in a low-resource language (denoted by $\mathcal{L}_{lr}$) can be further harnessed to explore similar content with minimal or no further training. Consider the illustrative examples from $\mathcal{D}_{hope}$ and $\mathcal{D}_{covid}$ below with loose translations:

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.

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.

Notice that both the low-resource language (Romanized Hindi) and world language (English) components exhibit similar intents of seeking peace (example from $\mathcal{D}_{hope}$) and urging people to comply with social distancing guidance (example from $\mathcal{D}_{covid}$). Our goal is to harness such low-resource components present in these highly code mixed documents to detect similar content written primarily in the low-resource language using the pipeline presented in Figure 2. Nearest neighbor sampling takes a seed set of comments and a larger pool of comments as inputs, and outputs comments from the larger pool that are nearest neighbors to the seed set. The distance metric is cosine-distance between the comment embeddings.

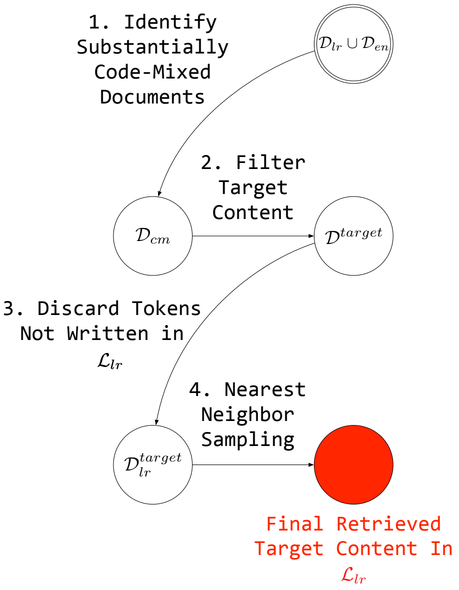
1. Identify Substantially Code-Mixed Documents
$\mathcal{D}_{lr} \cup \mathcal{D}_{en}$
$\mathcal{D}_{cm}$
2. Filter Target Content
$\mathcal{D}^{target}$
3. Discard Tokens Not Written in $\mathcal{L}_{lr}$
$\mathcal{D}_{lr}^{target}$
4. Nearest Neighbor Sampling
Final Retrieved Target Content In $\mathcal{L}_{lr}$

Figure 2: Target architecture. $\mathcal{L}_{lr}$ denotes a low-resource language.

Detecting code mixed documents: In order to detect highly code-mixed documents, we first need a token-level language identification tool. We found that $\widehat{\mathcal{L}}_{polyglot}$ is capable of performing token-level language identification without any modification (see, Table 2).

		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 $\mathcal{D}_{hope}$; en and h_e denotes English and Romanized Hindi, respectively.

Transfer in presence of a content classifier: In the *hope speech* detection task, we have access to a *hope speech* classifier [Palakodety *et al.*, 2020a] trained on English content. Hence, performing the steps presented in Figure 2 is straightforward. Highly code switched documents are detected using $\widehat{\mathcal{L}}_{polyglot}$; code-mixed *hope speech* is filtered out using the classifier; and then the Hindi parts of the *hope speech* comments are extracted by $\widehat{\mathcal{L}}_{polyglot}$; finally, nearest neighbor sampling is conducted. Table 3 lists example comments output by our cross-lingual sampling technique. Using our method, we obtain more than 10-fold improvement over our baseline (random sampling yields 1.8% *hope speech*).

Transfer in absence of a content classifier: On $\mathcal{D}_{covid}$, we address an extreme challenge of performing cross-lingual sampling when no content classifier exists, all we have is a handful of example English comments authored by annotators. Using a trick of selective deletion of English and Hindi parts, in [KhudaBukhsh *et al.*, 2020], we present a method that can circumvent the requirement of any content classifier. Even in this extreme setting, we obtained a 5-fold improvement over a random baseline. Example comments are listed in Table 4.

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 3: Random sample of *hope speech* obtained using our pipeline described in Figure 2.

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 4: Random sample of comments obtained through our method.

4 Lessons Learnt

Our main takeaways from this work are the following:

● *YouTube comments as a text data source:* In this work, and in our other works [Palakodety *et al.*, 2020b; Palakodety *et al.*, 2020c], we found that YouTube provides a rich alternative to Twitter in problems related to South Asia. It is the most popular social media platform in India with 265 million monthly active users (225 million on mobile), accounting for 80% of the population with internet access [HindustanTimes, 2019; YourStory, 2018]. Unlike Twitter, YouTube does not present any character limit per comment and hence allows substantially more elaborate and linguistically rich content than Twitter. However, YouTube has its fair share of disadvantages: unlike Twitter, the publicly available YouTube API does not provide geolocation information of a tweet, hence Twitter will be more appropriate for analyses where geolocation information is vital.

● *The non-native speaker aspect:* When dealing with content generated in a linguistically diverse region where the vast majority of the content contributors are non-native speakers in English, methods require to be substantially robust to cope with spelling and grammar disfluencies. For example, 32% of the times the word `liar` was mispelled as `lier` in $\mathcal{D}_{covid}$; our corpus contains several challenging examples as the following one: [`thankyou pakusta for hiumaniti no war aman ssnti kayam kare`] loosely translates to *Thank you Pakistan for humanity; let peace prevail*. Similar to the recent work in Pidgin English [Chang *et al.*, 2020], building robust methods that are aware of the quarks of English written by non-native speakers is an important research direction.

● $\widehat{\mathcal{L}}_{polyglot}$: Since viewers are commenting on videos relevant to a given issue, YouTube video comments are an effective way to gather highly topical discussions on a given matter. However, without efficient linguistic separation, using YouTube video comments as data source in a linguistically diverse region is not possible. Hence, in our projects, we found that $\widehat{\mathcal{L}}_{polyglot}$ was a critical component. We have made the code and implementation of $\widehat{\mathcal{L}}_{polyglot}$ publicly available to facilitate social science research[4].

● *Generality of our cross-lingual sampling technique:* On both domains of detecting *hope speech* and content encouraging COVID-19 health compliance, we observe that our technique generalizes well. We have not conducted any experiments on negative content (e.g., hate speech), but we believe our technique can be equally applicable in performing cross-lingual sampling of hate speech. Our finding that our technique survives even in absence of a content classifier or large number of labeled examples is particularly encouraging. In rapidly evolving scenarios (e.g., the current COVID-19 crisis where health guidelines are regularly updated), where class-definitions are changing, effective classifiers would require continual iteration. The fact that our sampling technique requires little resource indicates that it can be useful in evolving conditions.

● *Widening the support for Romanized Hindi:* In Indian social media, a substantial proportion of the corpus is non-English. In both of our data sets, comments written in Romanized Hindi outnumber comments written in English. Hindi has nearly 450 million speakers in India and Pakistan and social media content in Hindi primarily uses Roman script which has little linguistic support. In our recent work, we have scratched the surface of cross-lingual transfer to Romanized Hindi harnessing code switching [KhudaBukhsh *et al.*, 2020]; however, much remains to be done.

References

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

[BBC, 2020] BBC. Why has india's assam erupted over an 'anti-muslim' law?, 2020. Online; accessed 12-May-2020.

[Benesch *et al.*, 2016] Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright.

[4] https://www.cs.cmu.edu/~akhudabu/HopeSpeech.html

Counterspeech on twitter: A field study. *A report for Public Safety Canada under the Kanishka Project*, 2016.

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

[Bergsma *et al.*, 2012] Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. Language identification for creating language-specific twitter collections. In *Proceedings of the second workshop on language in social media*, pages 65–74. Association for Computational Linguistics, 2012.

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[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.

[Chang *et al.*, 2020] Ernie Chang, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. Unsupervised pidgin text generation by pivoting english data and self-training. *arXiv preprint arXiv:2003.08272*, 2020.

[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.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019.

[Facebook, 2016] Facebook. fastText: Language identification, 2016. [Online; accessed 5-June-2020].

[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, 2014.

[Google, 2020] Google. GoogleLangID, 2020. [Online; accessed 5-June-2020].

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

[Jaech *et al.*, 2016] Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. Hierarchical character-word models for language identification. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 84–93, Austin, TX, USA, November 2016. Association for Computational Linguistics.

[Jaidka *et al.*, 2019] Kokil Jaidka, Saifuddin Ahmed, Marko Skoric, and Martin Hilbert. Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, 29(3):252–273, 2019.

[Jauhiainen *et al.*, 2019] Tommi Sakari Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782, 2019.

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

[Joshi *et al.*, 2019] Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv preprint arXiv:1912.03457*, 2019.

[Kagan *et al.*, 2015] Vadim Kagan, Andrew Stevens, and VS Subrahmanian. Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election. *IEEE Intelligent Systems*, 30(1):2–5, 2015.

[KhudaBukhsh *et al.*, 2020] Ashiqur R. KhudaBukhsh, Shriphani Palakodety, and Jaime G. Carbonell. Harnessing code switching to transcend the linguistic barrier. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020 [scheduled for July 2020, Yokohama, Japan, postponed due to the Corona pandemic]*, pages 4366–4374, 2020.

[Koutra *et al.*, 2015] Danai Koutra, Paul N Bennett, and Eric Horvitz. Events and controversies: Influences of a shocking news event on information seeking. In *Proceedings of the 24th international conference on World Wide Web*, pages 614–624, 2015.

[Lui and Baldwin, 2014] Marco Lui and Timothy Baldwin. Accurate language identification of twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25, 2014.

[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.

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

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mulcaire *et al.*, 2018] Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith. Polyglot semantic role labeling. In *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 667–672, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Mulcaire *et al.*, 2019a] Phoebe Mulcaire, Jungo Kasai, and Noah A Smith. Low-resource parsing with crosslingual contextualized representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 304–315, 2019.

[Mulcaire *et al.*, 2019b] Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of NAACL:HLT*, pages 3912–3918, June 2019.

[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 to detect comments supporting the rohingyas. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 454–462, 2020.

[Roy *et al.*, 2017] Kaustav Roy, Disha Kohli, Rakeshkumar Kathirvel Senthil Kumar, Rupaksh Sahgal, and Wen-Bin Yu. Sentiment analysis of twitter data for demonetization in india–a text mining approach. *Issues in Information Systems*, 18(4):9–15, 2017.

[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.

[YourStory, 2018] YourStory. Youtube monthly user base touches 265 million in india, reaches 80 pc of internet population, 2018. Online; accessed 20-April-2020.