The Non-native Speaker Aspect: *Indian English* in Social Media

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

Abstract

As the largest institutionalized second language variety of English, *Indian English* has received a sustained focus from linguists for decades. However, to the best of our knowledge, no prior study has contrasted web-expressions of *Indian English* in noisy social media with English generated by a social media user base that are predominantly native speakers. In this paper, we address this gap in the literature through conducting a comprehensive analysis considering multiple structural and semantic aspects. In addition, we propose a novel application of language models to perform automatic linguistic quality assessment.

1 Introduction

Analyzing important issues through the lens of social media is a thriving field in computational social science (CSS) research. From policy debates (Demszky et al., 2019) to modern conflicts (Palakodety et al., 2020a), web-scale analyses of social media content present an opportunity to aggregate and analyze opinions at a massive scale. English being one of the widely-spoken pluricentric languages (Leitner, 1992), a considerable fraction of current CSS research primarily analyzes content authored in English. Several recent lines of CSS research on Indian sub-continental issues (Palakodety et al., 2020a; Tyagi et al., 2020; Palakodety et al., 2020c) dealt with *Indian English* (Mehrotra, 1998), a regional variant of English spoken in India and among the Indian diaspora.

As the largest institutionalized second language variety of English, *Indian English* has received sustained attention from linguists (Kachru, 1965; Shastri, 1996; Gramley and Pätzold, 2004; Sedlatschek, 2009) delineating multiple aspects in which Indian English is distinct from US or British English. However, these studies are largely confined to well-formed English written in formal settings (e.g., newspaper Dubey, 1989; Sedlatschek, 2009. The efforts so far in characterizing web-expressions of *Indian English* are somewhat scattered with isolated focus areas (e.g., code switching (Gella et al., 2014; Rudra et al., 2019; Khanuja et al., 2020; KhudaBukhsh et al., 2020a), use of swear words Agarwal et al., 2017, and word usage Kulkarni et al., 2016) and little attention given to analyzing the range of spelling, grammar and structural characteristics observed in web-scale *Indian English* corpora. Due to the deep penetration of cellphone technologies into Indian society and availability of inexpensive data (HuffPost, 2017), a user base with a wide range of English proficiency has access to the social media. Hence, understanding to what extent spelling and grammar issues affect *Indian English* found on the social web and how does that compare and contrast with typical noisy social media content generated by predominantly native English speakers is an important yet underexplored research question.

In this paper, via two substantial contemporaneous corpora constructed from comments on YouTube videos from major news networks in the US and India, we address the above research question. To the best of our knowledge, no prior study has contrasted any *Indian English* social media corpus with the variety of English observed in social media platforms frequented by native English speakers. We further use two existing corpora of news articles from India and the US to demonstrate that while college-educated, well-formed English across these two language centers does not differ by

*Ashiqur R. KhudaBukhsh is the corresponding author.

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 61–70

much, social media *Indian English* is different from social media *US English* on certain aspects, hence may pose a greater challenge to conduct meaningful analysis. Apart from using standard tools to assess linguistic quality, we present a novel finding that recent advances in language models can be leveraged to perform automated linguistic quality assessment of human-generated text.

2 Data Sets

We consider two social media (denoted by the superscript *sm*) data sets and two news article (denoted by the superscript *na*) data sets. We denote *Indian English* and *US English* as *en-in* and *en-us*, respectively. In order to keep our vocabulary statistics comparable, we sub-sample from our *en-us* social media data set and ensure that both social media corpora have nearly equal number of tokens. Detailed description of preprocessing steps are presented in the Appendix.

Why YouTube? Both of our social media corpora are comments on YouTube videos posted within an identical date range (30^{th} January, 2020 to 7^{th} May, 2020). As of January 2020, YouTube is the second-most popular social media platform in the world drawing 2 billion active users (Statista, 2020b). It is the most popular social media platform in India with 265 million monthly active users, accounting for 80% of the population with internet access (HindustanTimes, 2019; YourStory, 2018).

- $\mathcal{D}^{sm}_{en\text{-}in}$: We consider a subset of a data set first introduced in (KhudaBukhsh et al., 2020b). The original data set consists of 4,511,355 comments by 1,359,638 users on 71,969 YouTube videos from fourteen Indian news outlets posted between 30^{th} January, 2020 and 7^{th} May, 2020. Next, language is detected using $\hat{\mathcal{L}}_{polyglot}$, a polyglot embedding based language identifier first proposed in (Palakodety et al., 2020a) and successfully used in other multi-lingual contexts (Palakodety et al., 2020c). This yields 1,352,698 English comments (23,124,682 tokens, 2,107,233 sentences). In order to minimize the effects of code switching (Gumperz, 1982; Myers-Scotton, 1993), only sentences with low CMI (code mixing index) (Das and Gambäck, 2014) are considered. We estimate CMI using the same method presented in KhudaBukhsh et al. 2020a and set a threshold of 0.1. Upon removal of code switched sentences, our final data set, $\mathcal{D}^{sm}_{en\text{-}in}$, consists of 1,923,292 sentences (20,591,213 tokens).
- $\mathcal{D}^{sm}_{en\text{-}us}$: We consider a subset of a data set first introduced in (KhudaBukhsh et al., 2020c). We first obtain 10,245,348 comments posted by 1,690,589 users[1] on 8,593 YouTube videos from three popular US news channels (Fox news, CNN and MSNBC) (Statista, 2020a) in the same time period. We subsampled the data to make the number of tokens comparable to that of $\mathcal{D}^{sm}_{en\text{-}in}$. This resulted in $\mathcal{D}^{sm}_{en\text{-}us}$ having 1,573,355 sentences (20,591,220 tokens).
- $\mathcal{D}^{na}_{en\text{-}in}$ consists of 398,960 sentences (9,016,255 tokens) from news articles that appeared in highly circulated Indian news outlets (e.g., The Quint, Hindustan Times, Deccan Herald) (Dai, 2017).
- $\mathcal{D}^{na}_{en\text{-}us}$ consists of 94,463 sentences (2,042,024 tokens) from news articles that appeared in highly circulated US news outlets (e.g., HuffPost, Washington post, New York Times).

3 Analysis

3.1 Vocabulary and Grammar

We conduct a detailed study comparing and contrasting $\mathcal{D}^{sm}_{en\text{-}in}$ and $\mathcal{D}^{sm}_{en\text{-}us}$. In what follows, we summarize our observations (see, Appendix for details).

- ***Vocabulary***: In the context of social media, US English exhibits a richer overlap with standard English dictionary as compared to Indian English.

Let $\mathcal{V}_{dict}$ denote the English vocabulary obtained from a standard English dictionary (Kelly, 2016)[2]. Let $\mathcal{V}^{sm}_{en\text{-}in}$ and $\mathcal{V}^{sm}_{en\text{-}us}$ denote the vocabularies of $\mathcal{D}^{sm}_{en\text{-}in}$ and $\mathcal{D}^{sm}_{en\text{-}us}$, respectively. We now compute the following overlaps: $|\mathcal{V}^{sm}_{en\text{-}us} \cap \mathcal{V}^{dict}| = 43,826$ and $|\mathcal{V}^{sm}_{en\text{-}in} \cap \mathcal{V}^{dict}| = 38,260$. Also, with a list of 6,000 important words for US SAT exam[3], we find that $\mathcal{V}^{sm}_{en\text{-}us}$ has considerably larger overlap (4,349 words) than $\mathcal{V}^{sm}_{en\text{-}in}$ (3,956 words).

- ***Spelling deviations***: Indian English exhibits larger spelling deviations as compared to US English. Phonetic spelling errors (i.e., spelling a word as it sounds) are common in Indian English. This observation aligns with (KhudaBukhsh et al., 2020a).
- ***Loanwords***: Borrowed words, also known as loanwords, are lexical items borrowed from a donor language (Holden, 1976; Calabrese and

[1] The Jaccard similarity between the two social media user bases of $\mathcal{D}^{sm}_{en\text{-}in}$ and $\mathcal{D}^{sm}_{en\text{-}us}$ is 0.01 indicating minimal overlap between the two user bases.

[2] We take the union of *en-us* and *en-gb*.

[3] https://satvocabulary.us/INDEX.ASP?CATEGORY=6000LIST

Wetzels, 2009; Van Coetsem, 2016) . For example, the English word avatar or yoga is borrowed from Hindi. We observe that loanwords (e.g., `sadhus`, `begum`, `burqa`, `imams` and `gully`) borrowed from Hindi heavily feature in Indian English.

- ***Article and pronoun usage***: Indian English uses considerably fewer articles and pronouns as compared to US English. Pronoun and article omissions in ESL (English as Second Language) are well-studied phenomena (Ferguson, 1975). Our observation also aligns with a previous field study (Agnihotri et al., 1984) that reported even college-educated Indians make substantial errors in article usage.
- ***Preposition usage***: Indian English uses considerably fewer prepositions as compared to US English (11.48% in *en-us* and 10.84% in *en-in*).
- ***Verb usage***: Indian English uses fewer verbs than US English. Of the different verb forms (see, Figure 1), Indian English uses the root form relatively more than US English indicating (possible) poorer understanding of subject-verb agreement and tense (later verified in Section 3.2).
- ***Sentence length***: We observe shorter sentences in Indian as compared to US English (average *en-in* sentence length: 10.71 ± 12.37; average *en-us* sentence length: 13.09 ± 20.17). We acknowledge that device variability may influence this observation.
- ***Sentence validity evaluated by a parser***: A standard parser evaluates fewer Indian English sentences as valid as compared to US English (see, Table 1). However, no such discrepancy was observed in news article English from both language centers.
- ***Constituency parser depth***: For a given sentence length, Indian English exhibits lesser average constituency parser tree depth (Joshi et al., 2018) indicating (possible) structural issues. Intuitively, length of a sentence is likely to be positively correlated with its structural complexity; a long sentence is likely to have more complex (and nested) sub-structures than a shorter one. A parser's ability to correctly identify such sub-structures depends on the sentence's syntactic correctness. To tease apart the relationship between sentence-length and constituency parser's depth, in Figure 2, we present the average tree depth for a given sentence length. We observe that between well-formed English, the difference is almost imperceptible. However, as the sentence length grows, the gap between tree depth obtained in social media *en-in* and the rest widens indicating possible structural issues. A few example long sentences with small parse-tree depth are presented in the Appendix.

- ***Generalizability across other native English variants***: Our results are consistent when compared against a British English (*en-gb*) social media corpus.

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
Valid sentences	96.93	96.61	83.88	88.30

Table 1: Percentage of sentences determined valid by a constituency parser (Joshi et al., 2018).

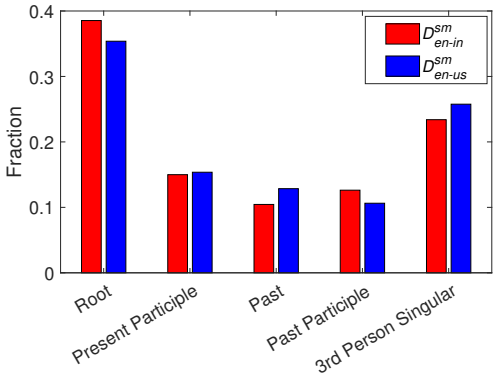

Figure 1: Distribution of different verb forms. We compute the relative occurrence of different morphological forms of a verb using a standard library (Honnibal and Montani, 2017).

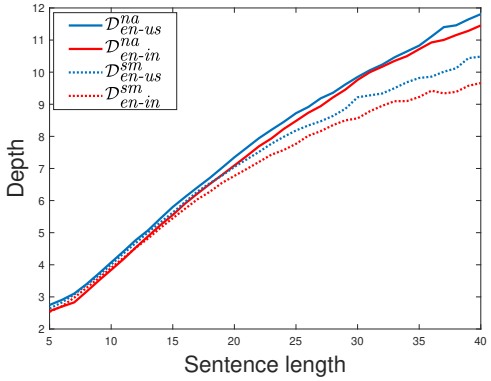

Figure 2: Constituency parser depth. A well-known parser (Joshi et al., 2018) is run on 10K sentences from each corpus. Average parse tree depth is presented for a given sentence length.

3.2 Cloze Test

Recent advances in Language Models (LMs) such as BERT (Devlin et al., 2019) have led to a substantial improvement in several downstream NLP tasks. While obtaining task-specific performance gain has been a key focus area (see, e.g., Liu and Lapata, 2019; Lee et al., 2020), several recent studies attempted to further the understanding of what exactly about language these models learn that results in these performance gains. LMs' ability to solve long-distance agreement problem and general syntactic abilities have been previously explored (Gulordava et al., 2018; Marvin and Linzen, 2018; Goldberg, 2019).

BERT's masked word prediction has a direct parallel in human psycholinguistics literature (Smith and Levy, 2011). When presented with a sentence (or a sentence stem) with a missing word, a cloze task is essentially a fill-in-the-blank task. For instance, in the following cloze task: *In the* `[MASK]`*, it snows a lot*, `winter` is a likely completion for the missing word. In fact, when given this cloze task to BERT, BERT outputs the following five seasons ranked by decreasing probability: `winter`, `summer`, `fall`, `spring` and `autumn`. Word prediction as a test of LM's language understanding has been explored in Paperno et al. (2016); Ettinger (2020) and recent studies leveraged it in novel applications such as relation extraction (Petroni et al., 2019) and political insight mining (Palakodety et al., 2020b). Bolstered by these findings and another recent result that uses BERT to evaluate the quality of translations (Zhang* et al., 2020), we propose an approach to estimate language quality using BERT. Our hypothesis is if a sentence is syntactically consistent and semantically coherent, BERT will be able to predict a masked word in that sentence with higher accuracy than a syntactically inconsistent or semantically incoherent sentence.

We first motivate our method with two examples. Consider the following classic syntactically correct yet semantically incoherent sentence (Chomsky, 1957): *Colorless green ideas sleep furiously*. BERT's top five predictions for a cloze task *Colorless green* `MASK` *sleep furiously* are the following: (1) `eyes` (2) `.` (3) `,` (4) `they` and (5) `I`. In fact, none of these words when masked, features in BERT's top 100 predictions. However, for another iconic sentence (King, 1968) when presented in the following cloze form: *I have a* `MASK` *that my four little children will one day live in a nation where they will not be judged by the color of their skin, but by the content of their character*, BERT's top five predictions (feeling, hope, belief, vision, and dream) correctly include *dream*. Notice that, in our semantically coherent example sentence, all of the predicted words have (Part-Of-Speech) POS agreement with the masked word while the semantically incoherent sentence produced a wide variety of completion choices that include punctuation, pronouns and noun.

We randomly sample 10k sentences from each corpus. For each sentence, we mask a randomly chosen word in the sentence such that $w \in \mathcal{V}_{dict}$ and construct an input cloze statement. Following standard practice (Petroni et al., 2019), we report p@1, p@5 and p@10 performance. p@K is defined as the top-K accuracy, i.e., an accurate completion of the masked word is present in the retrieved top K words ranked by probability.

Table 2 summarizes BERT's performance in predicting masked words. We were surprised to notice that on well-formed sentences, BERT achieved higher than 80% accuracy indicating that well-formed sentences leave enough cues for an LM to predict a masked word with high accuracy. We further observe that prediction accuracy is possibly correlated with linguistic quality; the performance on well-formed text corpora is substantially better than that of on social media text corpora. Finally, among the two social media text corpora, the performance on Indian English is substantially worse possibly indicating larger prevalence of grammar, spelling or semantic disfluencies. A few randomly sampled examples are listed in Table 4.

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
p@1	53.53	55.53	33.49	42.31
p@5	75.69	77.30	55.91	65.07
p@10	81.03	82.93	62.69	71.36

Table 2: BERT's masked word prediction performance on 10k randomly sampled sentences from each corpus. For each sentence, a word belonging to a standard English dictionary is randomly chosen for masking. Appendix contains additional results on a social media corpus of British English.

We next compute the fraction of instances where the POS tags of the masked word and the predicted word agree. As shown in Table 3, the POS agreements on well-formed text corpora are substantially higher than that of on the social media corpora.

Once again, we observe that of the two social media corpora, POS agreement on $\mathcal{D}^{sm}_{en\text{-}in}$ corpus is lower than that of on $\mathcal{D}^{sm}_{en\text{-}us}$. Our results inform that masked word prediction accuracy can be an effective measure in evaluating linguistic quality. Additional results with a British English corpus is presented in the Appendix.

	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
Overall	84.27	83.68	66.56	71.72
VERB	90.17	89.72	79.47	82.76
NOUN	86.20	85.95	62.03	67.96
ADP	89.78	89.24	75.96	75.88
ADJ	68.88	70.54	48.55	61.06
ADV	74.40	73.06	47.09	58.01

Table 3: POS agreement between the masked word and BERT's top prediction. Results are computed on 10K randomly sampled sentences from each corpus. Results on social media corpora are highlighted with blue. Adposition (ADP) is a cover term for prepositions and postpositions. ADJ and ADV denote adjective and adverb, respectively. Appendix contains additional results on a social media corpus of British English.

4 Conclusions

In this paper, we present a comprehensive comparative analysis between Indian English and US English in social media. Our analyses reveal that compared to native English, social media Indian English exhibits certain differences that may add to the challenges of navigating noisy, social media texts generated in the Indian sub-continent and thus present an opportunity to the NLP community to address these challenges. Recent lines of computational social science (CSS) research focusing on Indian sub-continental issues emphasized on the challenges faced while processing Indian social media data. However, no prior work contrasted social media *Indian English* with social media native English. We believe our work will help the research community identify focus areas to facilitate CSS research in this domain. We present a novel approach to perform automated linguistic quality assessment using BERT, a well-known high-performance language model. To the best of our knowledge, our work first tests BERT's masked word prediction accuracy on human-generated texts obtained from noisy social media. World variants of English spoken and written form have been widely studied for several decades. However, limited literature exists on characterizing their social media expressions.

Error type	Comment
SVD	The goons needs to be severely punished.
SVD	They play victim card, like they too suffering from virus.
SVD	every dogs come thier own day, you get what u deserves.
IVF	I am live in Assam.
IVF	these people will be never change.

Table 4: Random sample of sentences in $\mathcal{D}^{sm}_{en\text{-}in}$ with grammatical issues. SVD denotes subject-verb disagreement. IVF denotes incorrect verb forms

Our study makes a small step towards characterizing a broad range of aspects of Indian English observed in social media.

References

Prabhat Agarwal, Ashish Sharma, Jeenu Grover, Mayank Sikka, Koustav Rudra, and Monojit Choudhury. 2017. I may talk in english but gaali toh hindi mein hi denge: A study of english-hindi code-switching and swearing pattern on social networks. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, pages 554–557. IEEE.

Rama Kant Agnihotri, Amrit Lal Khanna, and Aditi Mukherjee. 1984. The use of articles in indian english: Errors and pedagogical implications. *IRAL: International Review of Applied Linguistics in Language Teaching*, 22(2):115.

Edward Loper Bird, Steven and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Andrea Calabrese and Leo Wetzels. 2009. *Loan phonology*. John Benjamins Publishing Company.

Noam Chomsky. 1957. Syntactic structures. the hague: Mouton.. 1965. aspects of the theory of syntax. *Cambridge, Mass.: MIT Press.(1981) Lectures on Government and Binding, Dordrecht: Foris.(1982) Some Concepts and Consequences of the Theory of Government and Binding. LI Monographs*, 6:1–52.

Tianru Dai. 2017. News Articles.

Amitava Das and Björn Gambäck. 2014. 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.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Vinod S Dubey. 1989. *Newspaper English in India*, volume 13. Bahri Publications.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguistics*, 8:34–48.

Charles A Ferguson. 1975. Toward a characterization of english foreigner talk. *Anthropological linguistics*, pages 1–14.

Spandana Gella, Kalika Bali, and Monojit Choudhury. 2014. "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.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Stephan Gramley and Kurt-Michael Pätzold. 2004. *A survey of modern English*. Routledge.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Kyril Holden. 1976. Assimilation rates of borrowings and phonological productivity. *Language*, pages 131–147.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

HuffPost. 2017. Youtube monthly user base touches 265 million in india, reaches 80 pc of internet population. Online; accessed 3-June-2020.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia. Association for Computational Linguistics.

Braj B Kachru. 1965. The indianness in indian english. *Word*, 21(3):391–410.

Ryan Kelly. 2016. Pyenchant a spellchecking library for python. *Ηλεκτρονικό]. Available: https://pythonhosted. org/pyenchant*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3575–3585. Association for Computational Linguistics.

Ashiqur R. KhudaBukhsh, Shriphani Palakodety, and Jaime G. Carbonell. 2020a. Harnessing code switching to transcend the linguistic barrier. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4366–4374. ijcai.org.

Ashiqur R. KhudaBukhsh, Shriphani Palakodety, and Tom M. Mitchell. 2020b. Discovering bilingual lexicons in polyglot word embeddings.

Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. 2020c. We don't speak the same language: Interpreting polarization through machine translation.

Martin Luther King. 1968. I have a dream. *Negro History Bulletin*, 31(5):16.

Vivek Kulkarni, Bryan Perozzi, Steven Skiena, et al. 2016. Freshman or fresher? quantifying the geographic variation of language in online social media. In *ICWSM*, pages 615–618.

J Lee, W Yoon, S Kim, D Kim, S Kim, CH So, and J Kang. 2020. Biobert: Pre-trained biomedical language representation model for biomedical text mining. arxiv 2019. *arXiv preprint arXiv:1901.08746*.

Gerhard Leitner. 1992. English as a pluricentric language. *Pluricentric languages: Differing norms in different nations*, 62:178–237.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Raja Ram Mehrotra. 1998. Indian english. *Texts and Interpretation. Amsterdam: Benjamins*.

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

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020a. Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1881–1889. IOS Press.

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

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020c. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 454–462.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Koustav Rudra, Ashish Sharma, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2019. Identifying and analyzing different aspects of english-hindi code-switching in twitter. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 18(3):29:1–29:28.

Andreas Sedlatschek. 2009. Contemporary indian english. *Variation and change. Amsterdam, Philadelphia*.

SV Shastri. 1996. Using computer corpora in the description of language with special reference to complementation in indian english. *South Asian English: structure, use, and users*, 2(4):70–81.

Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Statista. 2020a. Leading cable news networks in the united states in april 2020, by number of primetime viewers. Online; accessed 3-June-2020.

Statista. 2020b. Most popular social networks worldwide as of january 2020, ranked by number of active users. Online; accessed 3-June-2020.

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. 2020. A computational analysis of polarization on indian and pakistani social media. *arXiv preprint arXiv:2005.09803*.

Frans Van Coetsem. 2016. *Loan phonology and the two transfer types in language contact*, volume 27. Walter de Gruyter GmbH & Co KG.

YourStory. 2018. Youtube monthly user base touches 265 million in india, reaches 80 pc of internet population. Online; accessed 3-June-2020.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

5 Appendix

5.1 Data Sets

Preprocessing: We apply the following standard preprocessing steps on the raw comments.

- We convert all comments to lowercase and remove all emojis and junk characters.
- We replace multiple occurrences of punctuation with a single occurrence. For example, `they got trapped!!!!!!!` is converted into `they got trapped!`.
- We use an off-the-shelf sentence tokenizer from NLTK (Bird and Klein, 2009) to break up the comments into sentences.

YouTube channels: Table 5 lists the Indian YouTube channels we considered for $\mathcal{D}^{sm}_{en\text{-}in}$.

IndiaTV, NDTV India, Republic World, The Times of India, Zee News, Aaj Tak, ABP NEWS, CNN-News18, News18 India, NDTV, TIMES NOW, India Today, The Economic Times, Hindustan Times

Table 5: National channels.

5.2 Vocabulary and Grammar

Observation: *In the context of social media, US English exhibits a richer overlap with standard English dictionary as compared to Indian English.*
Analysis: Let $\mathcal{V}_{dict}$ denote the English vocabulary obtained from a standard English dictionary (Kelly, 2016)[4]. Let $\mathcal{V}^{sm}_{en\text{-}in}$ and $\mathcal{V}^{sm}_{en\text{-}us}$ denote the vocabularies of $\mathcal{D}^{sm}_{en\text{-}in}$ and $\mathcal{D}^{sm}_{en\text{-}us}$, respectively. We now compute the following overlaps: $|\mathcal{V}^{sm}_{en\text{-}us} \cap \mathcal{V}^{dict}| = 43,826$ and $|\mathcal{V}^{sm}_{en\text{-}in} \cap \mathcal{V}^{dict}| = 38,260$. Also, with a list of 6,000 important words for US SAT exam[5], $\mathcal{V}^{sm}_{en\text{-}us}$ has considerably larger overlap (4,349 words) than $\mathcal{V}^{sm}_{en\text{-}in}$ (3,956 words).

Observation: *In the context of social media, Indian English exhibits larger deviation from standard spellings as compared to US English.*
Analysis: We compute the extent of spelling deviations in the following way. For each out-of-vocabulary (OOV) word that has appeared at least 5 or more times in a given corpus, we use a standard spell-checker [6] to map it to a dictionary word present that also has appeared 5 or more times in the corpus. We observe that, overall, 9,653 *en-in* words had at least one or more spelling variations (or errors) while 5,436 *en-us* words had at least one or more spelling variations (or errors). The average number of variations (or errors) per word are 2.15 and 1.42 for *en-in* and *en-us*, respectively, indicating that Indian English exhibit larger deviation from standard spellings. Qualitatively, we notice that words with a large number of vowels are particularly prone to spelling variations (or errors), for instance, the word `violence` has the following misspellings in *en-in*: `voilence`, `voilance`, and `violance`. In *en-us*, `violence` did not have any high-frequent (occurring 5 or more times in the corpus) misspelling. We further observe that phonetic spelling errors are highly common in *en-in*. For instance, the word `liar` is often misspelled as `lier` and the word `people` is often misspelled as `peaple`.

Observation: *Loanwords borrowed from Hindi heavily feature in Indian English.*
Analysis: Table 6 lists highly frequent words that belong to one social media corpus but absent in the other. We observe that loanwords (Holden, 1976; Calabrese and Wetzels, 2009; Van Coetsem, 2016) (e.g., `sadhus`, `begum`, `burqa`, `imams` and `gully`) feature in Indian English. Few nouns are actually used in different proper noun contexts. For example, `raga`, originally a Hindi loanword that means a musical construct, is actually used to refer to **Ra**hul **Ga**ndhi, a famous Indian politician. Similarly, `newt` (a salamander species) and `tapper` refer to American politician Newt Gingrich and American journalist Jake Tapper, respectively. We note that terms specific to US politics (e.g., `gerrymandering`, `caucuses`, `senates`) and specific Indian political discourse (e.g., `demonetization`, `secularists`) solely appear in the relevant corpus. Words specific to Indian sports culture (e.g., `cricketers`) only appear in Indian English while US healthcare-specific words (e.g., `deductibles`) never appear in Indian English.

Observation: *Indian English uses considerably fewer articles and pronouns as compared to US English.*
Analysis: We next compute the respective unigram distributions $\mathcal{P}_{en\text{-}in}$ and $\mathcal{P}_{en\text{-}us}$. For each token

[4] We take the union of *en-us* and *en-gb*.
[5] https://satvocabulary.us/INDEX.ASP?CATEGORY=6000LIST
[6] https://norvig.com/spell-correct.html

Solely present in $\mathcal{V}^{sm}_{en\text{-}in}$	Solely present in $\mathcal{V}^{sm}_{en\text{-}us}$
sadhus, pelting, raga, begum, bole, indigo, demonetization, defaulters, bade, burqa, secularists, demonetisation, rioter, labourer, madrasas, rickshaw, gully, introspect, cricketers, defaulter, imams	tapper, impeachable, newt, caucuses, electable, subpoenas, jurors, mittens, clapper, brokered, reassigned, munchkin, gaffe, buybacks, senates, gerrymandering, impeachments, felonies, blowhard, centrists, deductibles

Table 6: Dictionary words solely present in one corpus but absent in the other corpus.

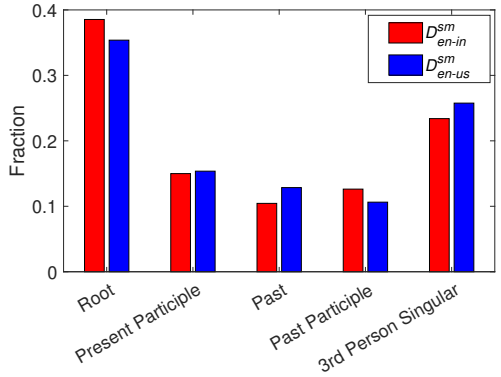

Figure 3: Distribution of different verb forms. We compute the relative occurrence of different morphological forms of a verb using a standard library (Honnibal and Montani, 2017).

$t \in \mathcal{V}_{dict} \cap \mathcal{V}^{sm}_{en\text{-}us} \cap \mathcal{V}^{sm}_{en\text{-}in}$, we compute the scores $\mathcal{P}_{en\text{-}in}(t) - \mathcal{P}_{en\text{-}us}(t)$, and $\mathcal{P}_{en\text{-}us}(t) - \mathcal{P}_{en\text{-}in}(t)$ and obtain the top tokens ranked by these scores (indicating increased usage in the respective corpus). Table 7 captures few examples with highest difference in unigram distribution. Overall, we notice that considerably fewer articles are used in *en-in*. Pronoun and article omissions in ESL (English as Second Language) are well-studied phenomena (Ferguson, 1975). This also aligns with a previous field study (Agnihotri et al., 1984) that reported even college-educated Indians make substantial errors in article usage.

Top tokens in $\mathcal{P}_{en\text{-}us}(t) - \mathcal{P}_{en\text{-}in}(t)$	Top tokens in $\mathcal{P}_{en\text{-}in}(t) - \mathcal{P}_{en\text{-}us}(t)$
the, a, trump, that, he, to, president, and, I, you, his, it, get, democrats, just, out, up, was, would, about	should, sir, u, police, good, are, in, govt, corona, very, please, is, these, them, congress, government, by, shame, only, pm

Table 7: Words with relatively more presence in one corpus over the other. Left column lists words that have relatively more presence in $\mathcal{D}^{sm}_{en\text{-}us}$ as compared to $\mathcal{D}^{sm}_{en\text{-}in}$ indicating that Indian English uses fewer articles and pronouns. Right column lists words that have relatively more presence in $\mathcal{D}^{sm}_{en\text{-}in}$ as compared to $\mathcal{D}^{sm}_{en\text{-}us}$.

Observation: *In the context of social media, Indian English uses considerably fewer prepositions as compared to US English.*

Analysis: We consider a list of highly frequent prepositions and find that *Indian English* uses fewer prepositions than US English (11.48% in *en-us* and 10.84% in *en-in*). We manually inspect usage of 100 randomly sampled sentences with the preposition `in`. 97 of such instances are evaluated correct by our annotators.

Observation: *In the context of social media, Indian English uses fewer verbs than US English.*

Analysis: In Figure 3, we summarize the relative occurrence of different verb forms. Of the different verb forms, Indian English uses the root form relatively more than US English indicating (possible) poorer understanding of subject-verb agreement and tense.

Observation: *In the context of social media, Indian English typically uses shorter sentences as compared to US English.*

Analysis: We use the recommended sentence tokenizer from NLTK (Bird and Klein, 2009) parser to obtain 1,923,292 and 1,573,355 sentences from $\mathcal{D}^{sm}_{en\text{-}in}$ and $\mathcal{D}^{sm}_{en\text{-}us}$, respectively. The average sentence length (by number of tokens) of $\mathcal{D}^{sm}_{en\text{-}in}$ and $\mathcal{D}^{sm}_{en\text{-}us}$ are 10.71 and 13.09, respectively. We acknowledge that device variability may influence this observation.

Observation: *A standard parser evaluates fewer Indian English sentences as valid as compared to US English.*

Analysis: We consider the same randomly sampled 10k sentences from each data set, and run a well-known constituency parser (Joshi et al., 2018). We first measure the fraction of sentences that are labeled as valid sentences by the parser. Table 8 shows that more than 96% sentences of both news article corpora are determined valid by the parser. Understandably, the fraction of valid sentences in the social media corpora is less with $\mathcal{D}^{sm}_{en\text{-}in}$ having few valid sentences than $\mathcal{D}^{sm}_{en\text{-}us}$.

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$
Valid sentences	96.93	96.61	83.88	88.30

Table 8: Percentage of sentences determined valid by a constituency parser (Joshi et al., 2018).

Observation: *For a given sentence length, Indian*

y another pm help fund what is the need of that coz there is already a pm relief fund and its has a committee with opposition party memeber too .
thanks to god that we have priminster like a modi ji he is our great mister i salute my priminister may god bless him always he has be long life
i request news reporter to use mask, plz do this, bcoz you are facing more dangerous situation only for public sake, plz sir i request to inform our reporter,they help usssssss
if sibal and singhvi becomes enjoy similar positions den its ok for dem kapilbsibal is ant national a gunda good for u sir dey r sour grapes and crook of sonia gandhi

Table 9: Random sample of long sentences in $\mathcal{D}^{sm}_{en\text{-}in}$ with low parse tree depth.

English exhibits lesser average constituency parser tree depth (Joshi et al., 2018) indicating (possible) structural issues.

Analysis: Intuitively, length of a sentence is likely to be positively correlated with its structural complexity; a long sentence is likely to have more complex (and nested) sub-structures than a shorter one. A parser's ability to correctly identify such sub-structures depends on the sentence's syntactic correctness. To tease apart the relationship between sentence-length and constituency parser's depth, in Figure 4, we present the average tree depth for a given sentence length. We observe that between well-formed English, the difference is almost imperceptible. However, as the sentence length grows, the gap between tree depth obtained in social media *en-in* and the rest widens indicating possible structural issues. A few example long sentences with small parse-tree depth are presented in Table 9.

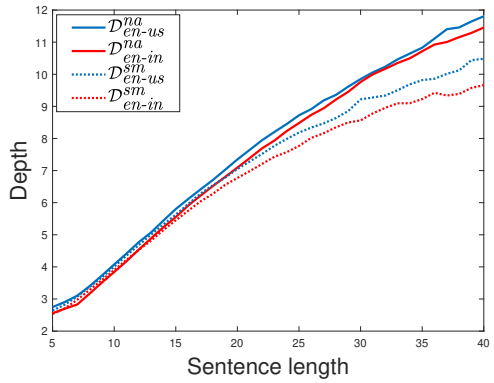

Figure 4: Constituency parser depth. A well-known parser is run on 10K sentences from each corpus. Average parse tree depth is presented for a given sentence length.

Observation: *Our results are consistent when compared against a British English (*en-gb*) social media corpus.*

Analysis: We construct an additional contemporaneous corpus of 4,034,513 comments from 57,019 videos from two highly popular British news outlets (BBC and Channel 4). In Table 10, 11 and 12, we present the results on British English and show that the results are very similar to US English.

	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}gb}$
Overall	84.27	83.68	66.56	71.72	71.41
VERB	90.17	89.72	79.47	82.76	78.30
NOUN	86.20	85.95	62.03	67.96	69.09
ADP	89.78	89.24	75.96	75.88	78.08
ADJ	68.88	70.54	48.55	61.06	58.17
ADV	74.40	73.06	47.09	58.01	62.25

Table 10: POS agreement between the masked word and BERT's top prediction. Results on social media corpora are highlighted with blue. Adposition (ADP) is a cover term for prepositions and postpositions.

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}gb}$
p@1	53.53	55.53	33.49	42.31	40.80
p@5	75.69	77.30	55.91	65.07	62.33
p@10	81.03	82.93	62.69	71.36	68.70

Table 11: BERT's masked word prediction performance on 10k randomly sampled sentences from each corpus.

Measure	$\mathcal{D}^{na}_{en\text{-}in}$	$\mathcal{D}^{na}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}in}$	$\mathcal{D}^{sm}_{en\text{-}us}$	$\mathcal{D}^{sm}_{en\text{-}gb}$
Valid sentences	96.93	96.61	83.88	88.30	84.17

Table 12: Percentage of sentences determined valid by a constituency parser.