Journal of Open
Humanities Data

Automatic Language Identification in Code-Switched Hindi-English Social Media Text

RESEARCH PAPER

]u[ubiquity press

LI NGUYEN
CHRISTOPHER BRYANT
SANA KIDWAI
THERESA BIBERAUER

*Author affiliations can be found in the back matter of this article

ABSTRACT

Natural Language Processing (NLP) tools typically struggle to process code-switched data and so linguists are commonly forced to annotate such data manually. As this data becomes more readily available, automatic tools are increasingly needed to help speed up the annotation process and improve consistency. Last year, such a toolkit was developed to semi-automatically annotate transcribed bilingual code-switched Vietnamese-English speech data with token-based language information and POS tags (hereafter the CanVEC toolkit, L. Nguyen & Bryant, 2020). In this work, we extend this methodology to another language pair, Hindi-English, to explore the extent to which we can standardise the automation process. Specifically, we applied the principles behind the CanVEC toolkit to data from the International Conference on Natural Language Processing (ICON) 2016 shared task, which consists of social media posts (Facebook, Twitter and WhatsApp) that have been annotated with language and POS tags (Molina et al., 2016). We used the ICON-2016 annotations as the gold-standard labels in the language identification task. Ultimately, our tool achieved an F_1 score of 87.99% on the ICON-2016 data. We then evaluated the first 500 tokens of each social media subset manually, and found almost 40% of all errors were caused entirely by problems with the gold-standard, i.e., our system was correct. It is thus likely that the overall accuracy of our system is higher than reported. This shows great potential for effectively automating the annotation of code-switched corpora, on different language combinations, and in different genres. We finally discuss some limitations of our approach and release our code and human evaluation together with this paper.

CORRESPONDING AUTHOR:
Li Nguyen
ALTA, Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
nhbn2@cam.ac.uk

KEYWORDS:
code-switching; language identification; automatic annotation; Hindi; English; Vietnamese

TO CITE THIS ARTICLE:
Nguyen, L., Bryant, C., Kidwai, S., & Biberauer, T. (2021). Automatic Language Identification in Code-Switched Hindi-English Social Media Text. *Journal of Open Humanities Data*, 7: 7, pp. 1–13. DOI: *https://doi.org/10.5334/johd.44*

1 CONTEXT AND MOTIVATION

In multilingual contexts, mixed output, featuring elements from two or more languages, is ubiquitous. Utterance (1), for example, demonstrates an instance of what is known as "code-switching", a construction in which a speaker alternates between different languages (in this case, Vietnamese and English).

(1) *mỗi* group *phải* *có* a different focus
each must have
"Each group must have a different focus."
(CanVEC, L. Nguyen & Bryant, 2020)

Although multilingualism is the norm world-wide (Grosjean & Li, 2013), NLP tools capable of processing more than one language per "sentential unit" as in (1) are still rather limited. This effectively circumscribes important applications such as machine translation (MT) and information retrieval (IR), and also the utility of NLP-based technology in contexts where language-users readily employ two or more languages side by side. Furthermore, as in other areas of NLP, while some efforts have been made to investigate somewhat high-resource language pairs such as English-Spanish (e.g. Ahn, Jimenez, Tsvetkov, & Black, 2020; Bullock, Guzmán, Serigos, Sharath, & Toribio, 2018; Solorio & Liu, 2008; Soto & Hirschberg, 2018) or English-Chinese (e.g. Chan, Ching, & Lee, 2005; Lyu, Dau-Cheng and Tan, Tien-Ping and Chng, Eng and Li, Haizhou, 2015; Shen, Wu, Yang, & Hsu, 2011), work examining code-switching involving low-resource, or less-described languages is still largely neglected. This means very few resources are available to automatically process this kind of data. With this in mind, two members of our team recently developed a toolkit to process the Canberra Vietnamese-English Corpus (CanVEC), an original corpus of 10 hours of natural mixed speech involving 45 Vietnamese-English migrant speakers living in Canberra. The corpus is semi-automatically annotated with language information and part-of-speech (POS) tags, obtaining >90% accuracy on both tasks (L. Nguyen & Bryant, 2020).

In this work, we test the wider feasibility of this framework in processing multilingual corpora by extending its application to another language pair, Hindi-English. Although Hindi-English is one of the more thoroughly investigated language pairs in the context of code-switching (e.g. Aguilar & Solorio, 2020; Bali, Sharma, Choudhury, & Vyas, 2014; Dey & Fung, 2014; Si, 2011), it nevertheless still provides a good test-bed in which to evaluate multilingual-corpus processing tools. We particularly focus on the language-identification task, for which we rely on the annotated data released in the International Conference on Natural Language Processing (ICON) 2016 shared task (Jamatia, Gambäck, & Das, 2015). In what follows, we report the result of this pilot as well as the challenges and implications that emerged.

2 RELATED WORK

It should be noted at the outset that language identification is one of the most important and well-studied tasks in computational approaches to code-switching. This is because it is often the prerequisite for many more complex downstream NLP tasks such as POS tagging, machine translation and speech recognition (Çetinoğlu, Schulz, & Vu, 2016; Choudhury, Chittaranjan, Gupta, & Das, 2014; Solorio & Liu, 2008). However, since monolingual processing tools tend to be less accurate in short or unidentified code-switching contexts, custom multi-lingual tools such as dictionary lookup, language models, morphological and phonological analysis, and machine learning techniques have become increasingly popular in recent years (Attia et al., 2019; Barman, Das, Wagner, & Foster, 2014; Mave, Maharjan, & Solorio, 2018; D. Nguyen & Doğruöz, 2013; Voss, Tratz, Laoudi, & Briesch, 2014; Xia, 2016). In particular, a wide range of machine learning algorithms such as Maximum Entropy, Naïve Bayes, Logistic Regression, and Support Vector Machines have been developed for code-switching language identification in many different language pairs or even triples, including English-Spanish (Mave et al., 2018; Solorio & Liu, 2008; Xia, 2016), English-Hindi (Mave et al., 2018), English-Mandarin (Lyu, Dau-Cheng and Tan, Tien-Ping and Chng, Eng and Li, Haizhou, 2015), Spanish-Wixarika (Mager, Çetinoğlu, & Kann, 2019), German-Turkish (Mager et al., 2019), Turkish-Dutch (D. Nguyen & Doğruöz, 2013), modern standard Arabic-Egyptian dialect (Elfardy, Al-Badrashiny, & Diab, 2013), English-Hindi-Bengali (Jamatia, Das, & Gambäck, 2019), and Romanized Moroccan Arabic (Darija)-English-French (Voss et al., 2014), among others. Performance is often reported to deliver a mid-90s F-score for

English-Spanish or English-Hindi, but much lower for less popular language pairs such as Arabic-Egyptian Arabic or Nepalese-English (80–85 F-score).

Machine learning methods, however, typically require a large amount of training data which may not always be available for low-resource languages participating in language contact. While this kind of data is nevertheless available for Hindi-English code-switching, in this study, we use it purely as a test set to investigate the performance of our approach in the hope that a similar methodology can also be applied to other, less well-resourced language pairs. We particularly hope that this will be of interest to traditional linguists who may be inexperienced with machine learning but who would otherwise have to annotate data manually.

3 METHODOLOGY

3.1 ICON-2016 DATA

The goal of the ICON-2016 shared task was to automatically annotate code-switched Hindi-English, Bengali-English and Telugu-English social media posts (Facebook/Twitter/WhatsApp) with either fine-grained or coarse-grained part-of-speech (POS) tags (Jamatia et al., 2015). Participants were provided with word tokenised[1] social media posts that were already annotated with native language information. Since the goal of this paper is to investigate the automatic annotation of language information in code-switched data, we ignore the POS annotations and only make use of the language tags. Specifically, we focus on the Hindi-English subset of the corpus for which there are seven possible tags (***Table 1***).

Table 1 The different language tags in the data, their meaning and some examples.

TAG	MEANING	EXAMPLE
en	English	I, the, songs, listening
hi	Hindi	*Apna* (mine), *ladki* (girl), *peeti* (drinks)
univ	Universal	#, !, @abc, #happy
mixed	Mixed	*Dedh*-litre (1.5 litre)
acro	Acronym	IITB, USA
ne	Named Entity	Europe, Paris
undef	Undefined	M

We downloaded the Facebook, Twitter and WhatsApp Hindi-English data from the shared task website.[2] The distribution of the seven language tags for each dataset and overall is shown in ***Table 2***.

Table 2 The distribution of language tags across datasets and overall.

TAG	FACEBOOK	TWITTER	WHATSAPP	OVERALL
en	13,214	3,732	363	17,309
hi	2,857	9,779	2,539	15,175
univ	3,628	3,354	281	7,263
mixed	7	1	0	8
acro	251	32	0	283
ne	656	413	35	1,104
undef	2	0	0	2
Total	20,615	17,311	3,218	41,144

[1] Not sentence segmented; i.e. each Facebook/Twitter/WhatsApp message may consist of more than one sentence or a single sentence may also be split across messages.

[2] Shared task website: *http://amitavadas.com/Code-Mixing.html*. Specifically, we downloaded the following 3 files:

Facebook: *http://amitavadas.com/ICON2016/FB_HI_EN_FN.txt*.
Twitter: *http://amitavadas.com/ICON2016/TWT_HI_EN_FN.txt*.
WhatsApp: *http://amitavadas.com/ICON2016/WA_HI_EN_FN.txt*.

Since several of these tags are relatively low frequency, we collapsed the **mixed, acro, ne** and **undef** tags into the **univ** category. This was partly because multi-class classification is more challenging with a greater number of labels (especially extremely rare labels), but also because we saw little reason to differentiate between these tags in the language identification task. For example, certain acronyms (e.g. DJ) and named entities (e.g. Holi) can be said to belong to both languages, yet are rarely indicative of code-switching. Similarly, while mixed tokens are certainly interesting examples of code-switching at a morphological level, they are extremely rare in the given dataset (N = 6) and so did not warrant a dedicated label.

The final distribution of labels across the reprocessed datasets is shown in ***Table 3***. It is interesting to note that the distribution of languages is different across datasets, with Facebook being predominantly English (64%), and Twitter and WhatsApp being predominantly Hindi (56% and 78% respectively). It is also notable that universal tokens comprise a significant proportion of the data and are roughly as prevalent as the minority code-switching language in all datasets.

Table 3 Final distribution of language tags after preprocessing.

TAG	FACEBOOK	TWITTER	WHATSAPP	OVERALL
en	13,214	3,732	363	17,309
hi	2,857	9,779	2,539	15,175
univ	4,544	3,800	316	8,660
Total	20,615	17,311	3,218	41,144

This can possibly be explained by the fact that social media data comes with its own set of particular challenges (as reviewed in Çetinoğlu et al., 2016), e.g. typos, intentional spelling deviations (e.g. "okkkk"), abbreviated Internet slang (e.g. "lol", "smh"), and non linguistic expressions (e.g. emoticons, URLs, hashtags, @ mentions, etc.), many of which are language-agnostic (i.e. universal). Universal tokens may thus be more prevalent in social media posts than other genres of text. These challenges nevertheless play a central role in our decision-making process, and will be discussed throughout this paper.

3.2 APPROACH

Following L. Nguyen and Bryant (2020), our approach to token-based language identification is rule-based and relies on a word list for each language. For English, we used a custom Hunspell word list that contained a combination of American, British, Canadian and Australian variant spellings.[3] It was important to allow all these variants in order to maximise the chance that a word would be properly classified. For Hindi, we used a list of 30,000 transliterations that had been extracted from an online Hindi lyric database (Gupta, Choudhury, & Bali, 2012) and made available in the Forum for Information Retrieval Evaluation (FIRE) 2013 shared task (Roy, Choudhury, Majumder, & Agarwal, 2013).[4] We used this dataset because social media users tend not to switch between Devanagari script for Hindi and Roman script for English, and instead use Roman script for everything, transliterating Hindi as necessary. Since there is no standard way of transliterating Hindi to English however (see Section 5 for more discussion), this list represents the largest resource we could find that also contains several variant Roman transliterations for the same Hindi word. We consequently hoped it would have sufficiently large coverage. It is worth mentioning that although an equivalent Hunspell word list for Hindi is also publicly available,[5] it uses Devanagari script and so is incompatible with the ICON-2016 data.

Before making use of these resources, however, we first wrote a number of rules to classify universal tokens that are language-agnostic. In particular, a token is classified as universal if it meets at least one of the following criteria:

[3] *http://wordlist.aspell.net/*.

[4] *http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/*.
We note that the dataset was recently removed from the shared task website but is still available via the Wayback Machine Internet archiving tool: *https://web.archive.org/web/20160312153954/http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/*.

[5] *https://github.com/Shreeshrii/hindi-hunspell*.

1. It does not contain any alphanumeric characters; e.g. punctuation;
2. It contains "@", "#" or "http", or else is "RT"; e.g. @usernames, #topics, URLs and retweets;
3. If non-alphanumeric characters are deleted, the string is a number; e.g. dates and times;
4. It starts with ":" or ";"; e.g. emoticons.

Having tagged universal tokens, the next step was to use the English and Hindi word lists. Specifically, if a token appears in the English word list, but not the Hindi word list, it is tagged as English, and if a token appears in the Hindi word list, but not the English word list, it is tagged as Hindi. This approach successfully accounted for the vast majority of tokens, but revealed 3,629 tokens that did not meet either criteria and were untagged. We hence extracted these tokens and annotated the top 1,000 most frequent ones manually. It is worth noting that 2,569 of the automatically untagged tokens only occurred once in the dataset, so we effectively only annotated tokens that appeared at least twice. The top 20 of these most frequent tokens and their counts are shown in ***Table 4***.

Table 4 The top 20 most frequent ambiguous-language tokens and their frequency.

TOKEN	FREQ.	TOKEN	FREQ.
to	556	this	134
I	496	my	126
a	357	for	126
of	258	aur	122
in	236	h	111
you	212	it	108
is	185	have	104
me	184	on	100
accha	152	or	91
ho	145	hi	88

Of the top 1,000 that were annotated, there were 59 tokens that we were unable to confidently classify as Hindi, English or universal. Most of these tokens (N = 41/59) were ambiguous high-frequency words in both languages; e.g. "to" which is a discourse marker in Hindi, and "me" which means either "I" or "in" in Hindi. Of the remaining unannotated tokens, three were unknown abbreviations ("clg", "mst", "em"), seven were mixed tokens from more than one language tag ("100ka", "sirji", "prajii", "newsAik", "masterni", "Ep3/18", "chahiyeShopkeeper"), and eight were simply unknown/indecipherable tokens ("o", "Yese", "furra", "fufa", "B", "t", "tem", "s").

Finally, whenever a token was not classified by any word list or rule, it was assigned a tag based on the previous non-universal token in the current message, or else tagged English if it was the first token in the sentence. The decision to ignore universal tokens in this manner was based on the observation that universal tokens form the rarest category and tend not to occur in long contiguous sequences, while the decision to use English as the default language for ambiguous first-word tokens was based solely on the observation that English is slightly more prevalent in the data than Hindi (17k vs. 15k tokens).[6] The final system hence classifies tokens according to the following ordered rules:

1. Assign label based on manually defined disambiguation word list; else;
2. Assign label based on universal token rules; else;
3. Assign label based on exclusive English or Hindi word list membership; else;
4. Assign label based on previous token label.

[6] Future work might prefer to label ambiguous first-word tokens according to the language of the following token rather than using a default.

It should be noted that the manual disambiguation list takes the highest priority in this system because manual human judgements are considered to be the most reliable.

4 EXPERIMENTS AND RESULTS

4.1 MANUAL DISAMBIGUATION LIST SIZE

We evaluated the effectiveness of our approach by comparing the predicted labels against the gold labels in terms of the F_1 score, which is a weighted average of *precision* (P) and *recall* (R). In particular, precision is calculated as the proportion of correct labels over predicted labels for a given tag (x_{cor}/x_{pred}), while recall is calculated as the proportion of correct labels over gold labels for a given tag (x_{cor}/x_{gold}). In other words, precision measures the extent to which a system can correctly predict a given tag (i.e. correctness), while recall measures the extent to which a system can correctly predict all intended instances of a given tag (i.e. coverage). The F-score is hence the harmonic mean of the two.[7] In the context of this work, we specifically compared the micro F_1 scores (which take the differences between class labels into account) using manual disambiguation lists of different sizes in order to better understand the relationship between manual annotation and performance; i.e. to what extent a larger word list increases performance. Results are shown in ***Figure 1***.

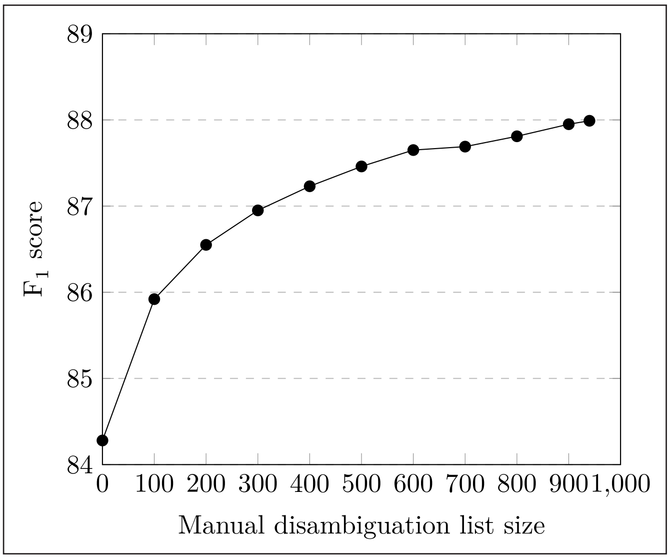

Figure 1 Language tagging performance as a function of manual disambiguation list size.

As expected, ***Figure 1*** shows diminishing returns as more manual labels are available. There is nevertheless a large gain from 84.2 to 86 F_1 for the first 100 manual tags, which shows that even a small word list of the most frequently ambiguous tokens can provide a significant boost to the overall performance. ***Figure 1*** also shows that this performance increase begins to level out at roughly 400–600 tokens, which roughly equates to tokens that occur at least 3–4 times or more in the data. This is a significant point to note as it potentially indicates an optimum level of manual annotation that should be carried out in future work (scaled according to the size of the data).

4.2 GENERAL EVALUATION

In addition to evaluating our system overall, we also evaluated in terms of P, R and F_1 for each language tag in each of the Facebook, Twitter and WhatsApp subsections of the overall corpus. The results are shown in ***Table 5*** where all systems make use of the full manual disambiguation list.

One of the most interesting results from this table is that performance on Hindi classification is stable across all datasets at 86–87 F_1, while performance on English classification varies considerably. Most notably, English classification scores almost 95.8 F_1 on the Facebook data, but just 53.1 F_1 on the WhatsApp data. This is largely due to precision being so low in the WhatsApp data (39.5). A similar effect is observed in the Twitter data, where the precision

[7] For more details on how F_1 score is computed, see e.g. Sasaki 2007.

TAG	FACEBOOK			TWITTER		
	P	R	F_1	P	R	F_1
en	93.34	98.35	95.78	70.26	81.32	75.39
hi	89.04	85.61	87.30	90.72	82.08	86.19
univ	97.36	84.51	90.48	80.35	87.61	83.82

TAG	WHATSAPP			OVERALL		
	P	R	F_1	P	R	F_1
en	39.52	80.99	53.12	85.98	94.32	89.95
hi	96.65	78.30	86.51	91.28	82.12	86.45
univ	59.71	78.80	67.94	87.23	85.66	86.44

Table 5 Precision, Recall and F_1 scores for each language tag in each corpus.

for English is the lowest out of the 3 tags at 70.3. Our first hypothesis for this observation was that the lower scores on the Twitter and WhatsApp data were a by-product of the decision to label unknown sentence-initial tokens as English by default. In particular, since the majority of tokens in the Twitter and WhatsApp data are Hindi, unlike the Facebook data, they would be more likely to benefit from Hindi as the default language. We hence tried labelling all unknown sentence-initial tokens (i.e. those that do not have a previous token) as Hindi rather than English, ultimately observing little improvement in the classification of English tokens in the Twitter data (75.4 F_1 → 76.5 F_1) and a noticeable improvement in the WhatsApp data (53.1 F_1 → 59.9 F_1). Precision in the WhatsApp data was nevertheless still very low at 39.5 → 49.8. In order to investigate why there might be such a difference between datasets and also to further evaluate the efficacy of our approach, we next carried out a manual evaluation of the first 500 tokens in each dataset.

4.3 QUALITATIVE EVALUATION

4.3.1 Coarse-grained

In our manual qualitative evaluation, we first annotated both the predicted and gold-standard language labels of the first 500 tokens in each dataset as either correct (COR) or incorrect (INC). While it might seem unusual to reannotate the gold standard for correctness, we encountered many cases where the gold standard was incorrect and we wanted to take this into account in the evaluation. *Table 6* hence shows the confusion matrices for all combinations of correct and incorrect labels in both our predictions (rows) and the gold standard (columns) for each dataset and overall.

FACEBOOK			TWITTER		
GOLD	COR	INC	GOLD	COR	INC
PRED			PRED		
COR	466	6	COR	425	25
INC	24	4	INC	35	15

WHATSAPP			OVERALL		
GOLD	COR	INC	GOLD	COR	INC
PRED			PRED		
COR	403	49	COR	1294	80
INC	41	7	INC	100	26

Table 6 Confusion matrices for correct (COR) and incorrect (INC) labels in each dataset.

This table shows that there were 1294/1500 (86%) tokens across all datasets where both the prediction and gold standard were correct. There were a further 80/1500 (5%) tokens where our prediction was correct but the gold standard was incorrect (49 of which occurred in the WhatsApp data), and 100/1500 (7%) tokens where our prediction was incorrect but the gold standard was correct. The remaining 26/1500 (2%) tokens were incorrect in both the prediction and gold standard. The most significant finding from these results is that of the 206/1500 tokens where at least one label was considered incorrect, just over half of them (106/206) were in the gold standard. This suggests our classifier may actually be more reliable than reported above, as almost 40% of all errors are caused by problems with the dataset. It is also notable that most of the gold-standard errors occurred in the WhatsApp and Twitter data, which suggest these datasets are noisier than the Facebook data. Examples of gold-standard errors include English abbreviations that were tagged as Hindi (e.g. “thnk u” (for “thank you”) and “ofc” (for “of course”)), universal emojis that were tagged as Hindi (e.g. “😁”), and real English words that were tagged as either Hindi or universal (e.g. “life” and “path”).

4.3.2 Fine-grained

To further investigate the limitations of our approach, we also manually classified the 126/1500 errors made by our system into five different categories depending on the perceived reason for the error. The definitions of the categories and examples are shown in *Table 7*.

Table 7 The five different types of classification errors with examples.

CODE	MEANING	EXAMPLES (AND INCORRECT PREDICTED TAG)		
A	Tokenisation/Orthography	^LøVĕ^ (hi)	-*Subha (en)	2014–15)ka (en)
B	Named entity	Tanzeel (en)	Amir (hi)	chennai (en)
C	Token in both word lists	he (en)	to (en)	are (en)
D	Token in neither word list	Achhi (en)	Namaskar (en)	tiket (hi)
E	Token in incorrect word list	Mt (en)	thy (en)	pre (hi)

More specifically, tokens were classified as Type A when the error was the result of incorrect tokenisation or non-standard orthography, Type B when the token was a named entity that was not classified as universal, Type C or D when either the token was a frequently-used word in both word lists or a rare token/spelling error in neither word list and it was furthermore incorrect to rely on the language of the previous token, and Type E when the token occurred only in the word list of the incorrect language. The results are shown in *Table 8*.

Table 8 The error type distribution between datasets.

CODE	FACEBOOK	TWITTER	WHATSAPP	OVERALL
A	3	20	1	24
B	5	16	7	28
C	4	4	21	29
D	12	8	12	32
E	4	2	7	13
			Total	126

One of the most significant findings from this table is that, overall, out of the few errors that our system failed to correct, no single category significantly outnumbered the rest. This suggests there is no obvious weakness to our classifier. We do note, however, that the distribution of error types can vary significantly between datasets. For example, Twitter has the highest incidence of Type A tokenisation errors (N = 20/24), while WhatsApp has the highest incidence of Type C ‘both word list’ errors (N = 21/29). On closer inspection, we found that the former was caused by a single tweet in the Twitter dataset that contained repeated multi-punctuation strings of the form “:-*Subha”, which were systematically tokenised incorrectly (N = 15/24 errors), while the latter was an artefact of shorter messages and slang in the WhatsApp dataset. Specifically, since WhatsApp messages are much shorter than Facebook or Twitter posts (typically only 2–5 words), this meant there was a lower chance for a token to have a reliable previous language context if it was ambiguous in both word lists. This unique property of the WhatsApp dataset is hence something to be aware of when processing very short messages.

In summary, we note that our approach is quite robust for processing Hindi-English social media code-switched text. This is significant because the methodology was originally developed to process transcribed natural code-switched speech between Vietnamese and English, an entirely different dataset both in terms of the languages involved and the media through which the code-switching is conducted. This highlights the potential for further extending the approach to different code-switched datasets across different media and language pairs.

5 NLP CHALLENGES IN PROCESSING MULTILINGUAL DISCOURSE

Despite this encouraging result, it is worth noting that several challenges in processing multilingual discourse remain. The first of these is specific to processing social media Hindi data. Specifically, Hindi is traditionally written in Devanagari script; however, social media users primarily use Roman script to write Hindi, in what is sometimes called Romanagari script (Bali et al., 2014; V.B., Choudhury, Bali, Dasgupta, & Basu, 2010; Virga & Khudanpur, 2003).[8] Although there are several commonly used conventions for Romanagari, there is no standardised spelling. For example, "d" is used for "द" /d̪/ (dental d), "ड" /ɖ/ (retroflex d), and sometimes "ड़" /ɽ/ (retroflex r). Many-to-one mappings in the Devanagari-Roman direction are also caused by dialectal differences at times. People tend to transliterate based on what they hear rather than formal Devanagari spellings. For example, "ज़" /z/ is pronounced as /ʤ/ in some dialects and so is represented as "z" or "j" in Roman script despite always being written as "ज़" in Devanagari. This, together with the fact that Hindi has a larger inventory of consonants and vowels (11 vowels and 35 consonants in Devanagari script[9] vs. 5 vowels and 21 consonants in the Roman script), highlights a lack of one-to-one mapping between Devanagari and Roman letters and leads to several issues in writing Romanagari (Mhaiskar, 2015).

The second problem, which remains challenging across the field is the inherent bias towards English (see e.g. Anastasopoulos & Neubig, 2020; Garrido-Muñoz, Montejo-Ráez, Martínez-Santiago, & Ureña-Lápez, 2021 for some recent overview), both in terms of available resources and human judgements. In our case, for example, most of the errors are target Hindi tokens. *Table 9* illustrates.

Table 9 Distribution of error types based on the target gold standard.

CODE	TYPE	TARGET				
		ENGLISH	HINDI	UNIVERSAL	UNDEFINED	OVERALL
A	Tokenisation/Orthography	1	18	2	3	24
B	Named entity	0	0	28	0	28
C	Token in both word lists	1	28	0	0	29
D	Token in neither word list	7	21	4	0	32
E	Token in incorrect word list	2	9	2	0	13
Total		11	76	35	3	126

It is clear from the table that the target Hindi errors significantly outnumber those of English and universal tokens (N = 76/126 compared to 11/126 and 35/126 respectively).[10] Although the Hindi word list we used was specifically chosen to offset the lack of standardised Romanagari spellings, in that it featured commonly used alternative spellings for each word, the high degree of variability in Romanagari spellings meant that some spelling possibilities were inevitably missing. These missing spellings led to a high number of Type D (no word list) errors for target Hindi words (N = 21/32). There were also some spelling alternatives that were missing in the Hindi word list but were found in the English word list instead (Type E 'incorrect word list' target Hindi errors N = 9/13). This is because the majority of these errors (N = 8/9) involved very short Hindi words with omitted vowels, which coincidentally constituted English abbreviations in the word list and were consequently incorrectly tagged as English (e.g. "mt" represents "mʉ" in

[8] The same holds for other Indian languages, such as Marathi (also traditionally written in Devanagari) (Mhaiskar, 2015) and Punjabi (traditionally written in Gurmukhi) (Kaur & Singh, 2015), as well as various dialects of modern Arabic (Eskander, Al-Badrashiny, Habash, & Rambow, 2014).

[9] There is disagreement on exact numbers. The numbers given are from the Government of India as reported by the BBC: *https://www.bbc.co.uk/languages/other/hindi/guide/alphabet.shtml*.

[10] Note that the **undefined** tokens were made up by 3/24 Type A errors that could not be attributed to any target tag as they were mixed language tokens, e.g. "Girl-Sacchi" [en-hi] and "haiAnother" [hi-en].

Hindi meaning “do not”, but is an abbreviation in English meaning “mountain”). These Hindi-specific issues are particularly amplified by social media text, which is self-transcribed by each speaker and so a single spelling convention is not used. We suggest normalisation of spelling and/or using a more comprehensive Hindi word list as a way to improve performance.

Furthermore, the bias towards English is not constrained solely by available resources but also extends to human judgements. For example, the dataset contained the words “India” and “Bharat” which are the English and Hindi names for the same named entity respectively. Although they should thus both be tagged as universal, we noted a preference by the annotators for tagging “India” as universal but “Bharat” as Hindi. Upon recognising this bias, we ultimately decided that both the language-specific tag (i.e. English for “India” and Hindi for “Bharat”) as well as the universal tag were equally valid answers. This example nevertheless shows that while English named entities are often more likely to be considered universal, perhaps partly due to the status of English as a global lingua franca, Hindi named entities may be more ambiguous, especially if they have an English counterpart. This possible bias is something that annotators should keep in mind for future work.

6 IMPLICATIONS

In this paper, we examined the extent to which we could standardise the automated processing of multilingual corpora, using a rule-based system originally developed to annotate transcribed bilingual code-switched Vietnamese-English speech data (L. Nguyen & Bryant, 2020). We applied this approach to Hindi-English social media text and achieved a high performance of 87.99 F_1 on the language identification task. We furthermore carried out an error analysis and found that almost 40% of all classification errors were caused by problems with the gold standard, and so performance is actually likely to be even higher. These findings are particularly promising given the inherently challenging nature of social media text as well as the idiosyncratic conventions of the language pairs involved.

In the broader context, our work further highlighted how well a rule-based system can handle various kinds of code-switched input. In particular, we found that the approach generalises to both isolating (i.e. Vietnamese) and fusional (i.e. Hindi) language pairs with English, and is not dependent on annotated training data for machine learning. Ultimately, the most significant challenge is to instead obtain a suitably diverse word list which is not just limited to standardised spellings. Unfortunately, however, research in multilingual NLP has rarely considered other languages that may not have standardised orthography, or whose varieties may not be so well-established. In an era where the worldwide ‘normality’ of multilingualism becomes increasingly visible and language innovation continues to speedily spread, this lack of resources poses an even more urgent problem. Devising an efficient way to create and update different word lists across different language varieties is thus a worthwhile avenue for future research.

ADDITIONAL FILES

The resources associated with this paper can be accessed at *https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QD94F9*.

FUNDING STATEMENT

We gratefully acknowledge the Cambridge Language Sciences Incubator Fund for supporting this project. At the time that this work was conducted, Nguyen was also funded by the Cambridge Trust International Scholarship, Bryant and Biberauer by Cambridge Assessment, and Kidwai by a Harding Distinguished Postgraduate Scholarship.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Li Nguyen: Conceptualisation, funding acquisition, project administration, resources, formal analysis, writing – original draft, writing – review & editing.

Christopher Bryant: Conceptualisation, funding acquisition, data curation, investigation, methodology, validation, writing – original draft.

Sana Kidwai: Data curation, formal analysis, methodology, resources, visualisation, writing – original draft.

Theresa Biberauer: Conceptualisation, funding acquisition, supervision, writing – review & editing.

AUTHOR AFFILIATIONS

Li Nguyen *orcid.org/0000-0001-8632-7909*
ALTA, Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
Christopher Bryant *orcid.org/0000-0002-1015-9467*
ALTA, Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
Sana Kidwai *orcid.org/0000-0003-4834-0541*
Section of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK
Theresa Biberauer *orcid.org/0000-0003-3840-7618*
Section of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK

REFERENCES

Aguilar, G., & **Solorio, T.** (2020, July). From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8033–8044). Online: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/2020.acl-main.716*. DOI: *https://doi.org/10.18653/v1/2020.acl-main.716*

Ahn, E., Jimenez, C., Tsvetkov, Y., & **Black, A. W.** (2020, January). What code-switching strategies are effective in dialogue systems? In *Proceedings of the Society for Computation in Linguistics 2020* (pp. 254–264). New York, USA: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/2020.scil-1.32*

Anastasopoulos, A., & **Neubig, G.** (2020, July). Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8658–8679). Online: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/2020.acl-main.766*. DOI: *https://doi.org/10.18653/v1/2020.acl-main.766*

Attia, M., Samih, Y., Elkahky, A., Mubarak, H., Abdelali, A., & **Darwish, K.** (2019, August). POS tagging for improving code-switching identification in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (pp. 18–29). Florence, Italy: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W19-4603*. DOI: *https://doi.org/10.18653/v1/W19-4603*

Bali, K., Sharma, J., Choudhury, M., & **Vyas, Y.** (2014, October). "I am borrowing ya mixing?"An analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 116–126). Doha, Qatar: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W14-3914*. DOI: *https://doi.org/10.3115/v1/W14-3914*

Barman, U., Das, A., Wagner, J., & **Foster, J.** (2014, October). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 13–23). Doha, Qatar: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W14-3902*. DOI: *https://doi.org/10.3115/v1/W14-3902*

Bullock, B., Guzmán, W., Serigos, J., Sharath, V., & **Toribio, A. J.** (2018, July). Predicting the presence of a Matrix Language in code-switching. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 68–75). Melbourne, Australia: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W18-3208*. DOI: *https://doi.org/10.18653/v1/W18-3208*

Çetinoğlu, Ö., Schulz, S., & **Vu, N. T.** (2016, November). Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 1–11). Austin, USA: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W16-5801*. DOI: *https://doi.org/10.18653/v1/W16-5801*

Chan, J. Y. C., Ching, P. C., & **Lee, T.** (2005). Development of a Cantonese-English code-mixing speech corpus. In *Proceedings of the Ninth European Conference on Speech Communication and Technology – Interspeech'05* (pp. 1533–1536). Lisbon, Portugal. Retrieved from *https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_1533.pdf*

Choudhury, M., Chittaranjan, G., Gupta, P., & **Das, A.** (2014). *Overview of FIRE 2014 Track on Transliterated Search* (Tech. Rep.). Retrieved from *https://www.isical.ac.in/~fire/working-notes/2014/MSR/2014-trainslit_search-track_over.pdf*

Dey, A., & **Fung, P.** (2014, May). A Hindi-English code-switching corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from *http://www.lrec-conf.org/proceedings/lrec2014/pdf/922_Paper.pdf*

Elfardy, H., Al-Badrashiny, M., & **Diab, M.** (2013). Codeswitch point detection in Arabic. In E. Métais, F. Meziane, M. Saraee, V. Sugumaran & S. Vadera (Eds.), *Natural Language Processing and Information Systems* (pp. 412–416). Berlin, Germany: Springer. DOI: *https://doi.org/10.1007/978-3-642-38824-8_51*

Eskander, R., Al-Badrashiny, M., Habash, N., & **Rambow, O.** (2014, October). Foreign words and the automatic processing of Arabic social media text written in Roman script. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 1–12). Doha, Qatar: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W14-3901*. DOI: *https://doi.org/10.3115/v1/W14-3901*

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., & **Ureña-Lápez, L. A.** (2021). A survey on bias in deep NLP. *Applied Sciences*, *11*(7), 3184. Retrieved from *https://www.mdpi.com/2076-3417/11/7/3184*. DOI: *https://doi.org/10.3390/app11073184*

Grosjean, F., & **Li, P.** (2013). *The psycholinguistics of bilingualism*. Chichester, UK: Wiley-Blackwell.

Gupta, K., Choudhury, M., & **Bali, K.** (2012, May). Mining Hindi-English transliteration pairs from online Hindi lyrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2459–2465). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from *http://www.lrec-conf.org/proceedings/lrec2012/pdf/365_Paper.pdf*

Jamatia, A., Das, A., & **Gambäck, B.** (2019). Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora. *Journal of Intelligent Systems*, *28*(3), 399–408. DOI: *https://doi.org/10.1515/jisys-2017-0440*

Jamatia, A., Gambäck, B., & **Das, A.** (2015, September). Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 239–248). Hissar, Bulgaria: INCOMA Ltd. Retrieved from *https://www.aclweb.org/anthology/R15-1033*

Kaur, J., & **Singh, J.** (2015). Toward normalizing Romanized Gurumukhi text from social media. *Indian Journal of Science and Technology*, *8*(27), 1–6. DOI: *https://doi.org/10.17485/ijst/2015/v8i27/81666*

Lyu, D-C., Tien-Ping, T., Eng, C., & **Haizhou, L.** (2015). Mandarin–English codeswitching speech corpus in South-East Asia: SEAME. *Language Resources and Evaluation*, *49*, 1986–1989. DOI: *https://doi.org/10.1007/s10579-015-9303-x*

Mager, M., Çetinoğlu, Ö., & **Kann, K.** (2019, June). Subword-level language identification for intraword code-switching. 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)* (pp. 2005–2011). Minneapolis, USA: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/N19-1201*. DOI: *https://doi.org/10.18653/v1/N19-1201*

Mave, D., Maharjan, S., & **Solorio, T.** (2018, July). Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 51–61). Melbourne, Australia: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W18-3206*. DOI: *https://doi.org/10.18653/v1/W18-3206*

Mhaiskar, R. (2015). Romanagari an alternative for modern media writings. *Bulletin of the Deccan College Research Institute*, *75*, 195–202. Retrieved from *http://www.jstor.org/stable/26264736*

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., & **Solorio, T.** (2016, November). Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 40–49). Austin, USA: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W16-5805*. DOI: *https://doi.org/10.18653/v1/W16-5805*

Nguyen, D., & **Doğruöz, A. S.** (2013, October). Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 857–862). Seattle, USA: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/D13-1084*

Nguyen, L., & **Bryant, C.** (2020, May). CanVEC – the Canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)* (pp. 4121–4129). Marseille, France: European Language Resources Association. Retrieved from *https://www.aclweb.org/anthology/2020.lrec-1.507*

Roy, R. S., Choudhury, M., Majumder, P., & **Agarwal, K.** (2013, December). Overview of the FIRE 2013 track on transliterated search. In *FIRE'12 & '13: Post-Proceedings of the Fourth and Fifth Workshops of the Forum for Information Retrieval Evaluation* (pp. 1–7). New York, USA: Association for Computing Machinery. DOI: *https://doi.org/10.1145/2701336.2701636*

Sasaki, Y. (2007). *The truth of the F-measure (Tech. Rep.)*. Manchester, UK: University of Manchester. Retrieved from *https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf*

Shen, H. P., Wu, C. H., Yang, Y. T., & **Hsu, C. S.** (2011, October). CECOS: A Chinese-English code-switching speech database. In *2011 International Conference on Speech Database and Assessments, Oriental COCOSDA 2011 – Proceedings* (pp. 120–123). Hsinchu City, Taiwan. DOI: *https://doi.org/10.1109/ICSDA.2011.6085992*

Si, A. (2011). A diachronic investigation of Hindi–English code-switching, using Bollywood film scripts. *International Journal of Bilingualism*, *15*(4), 388–407. DOI: *https://doi.org/10.1177/1367006910379300*

Solorio, T., & **Liu, Y.** (2008, October). Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 1051–1060). Honolulu, Hawaii: Association for Computational Linguistics. Retrieved from *http://aclweb.org/anthology/D08-1110*. DOI: *https://doi.org/10.3115/1613715.1613852*

Soto, V., & **Hirschberg, J.** (2018, July). Joint part-of-speech and language ID tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 1–10). Melbourne, Australia: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W18-3201*. DOI: *https://doi.org/10.18653/v1/W18-3201*

Sowmya, V. B., Choudhury, M., Bali, K., Dasgupta, T., & **Basu, A.** (2010, May). Resource creation for training and testing of transliteration systems for Indian languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). Retrieved from *http://www.lrec-conf.org/proceedings/lrec2010/pdf/182_Paper.pdf*

Virga, P., & **Khudanpur, S.** (2003). Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition – Volume 15* (pp. 57–64). Sapporo, Japan: Association for Computational Linguistics. DOI: *https://doi.org/10.3115/1119384.1119392*

Voss, C., Tratz, S., Laoudi, J., & **Briesch, D.** (2014, May). Finding Romanized Arabic dialect in code-mixed Tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2249–2253). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from *http://www.lrec-conf.org/proceedings/lrec2014/pdf/1116_Paper.pdf*

Xia, M. X. (2016, November). Codeswitching language identification using subword information enriched word vectors. In *Proceedings of the second workshop on computational approaches to code switching* (pp. 132–136). Austin, USA: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/W16-5818*. DOI: *https://doi.org/10.18653/v1/W16-5818*

TO CITE THIS ARTICLE:
Nguyen, L., Bryant, C., Kidwai, S., & Biberauer, T. (2021). Automatic Language Identification in Code-Switched Hindi-English Social Media Text. *Journal of Open Humanities Data*, 7: 7, pp. 1–13. DOI: *https://doi.org/10.5334/johd.44*

Published: 25 June 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.