A Twitter Corpus for Hindi-English Code Mixed POS Tagging

Kushagra Singh
IIIT Delhi

Indira Sen
IIIT Delhi

Ponnurangam Kumaraguru
IIIT Delhi

{kushagra14056,indira15021,pk}@iiitd.ac.in

Abstract

Code-mixing is a linguistic phenomenon where multiple languages are used in the same occurrence that is increasingly common in multilingual societies. Code-mixed content on social media is also on the rise, prompting the need for tools to automatically understand such content. Automatic Parts-of-Speech (POS) tagging is an essential step in any Natural Language Processing (NLP) pipeline, but there is a lack of annotated data to train such models. In this work, we present a unique language tagged and POS-tagged dataset of code-mixed English-Hindi tweets related to five incidents in India that led to a lot of Twitter activity. Our dataset is unique in two dimensions: **(i)** it is larger than previous annotated datasets and **(ii)** it closely resembles typical real-world tweets. Additionally, we present a POS tagging model that is trained on this dataset to provide an example of how this dataset can be used. The model also shows the efficacy of our dataset in enabling the creation of code-mixed social media POS taggers.

1 Introduction

With the rise of Web 2.0, the volume of text on Online Social Networks (OSN) has grown. Bilingual or trilingual social media users have thus contributed to a multilingual corpus containing a combination of formal and informal posts. Code-switching or code-mixing[1] occurs when ”lexical items and grammatical features from two languages appear in one sentence” (Muysken, 2000). It is frequently seen in multilingual communities and is of interest to linguists due to its complex relationship with societal factors. Past research has looked at multiple dimensions of this behaviour, such as it’s relationship to emotion expression (Rudra et al., 2016) and identity. But research efforts are often hindered by the lack of automated Natural Language Processing (NLP) tools to analyze massive amounts of code-mixed data (Bali et al., 2014). POS tags are used as features for downstream NLP tasks and past research has investigated how to obtain accurate POS tags for noisy OSN data. POS tagging for Code-mixed social media data has also been investigated (Gimpel et al., 2011), however, existing datasets are either hard to obtain or lacking in comprehensiveness.

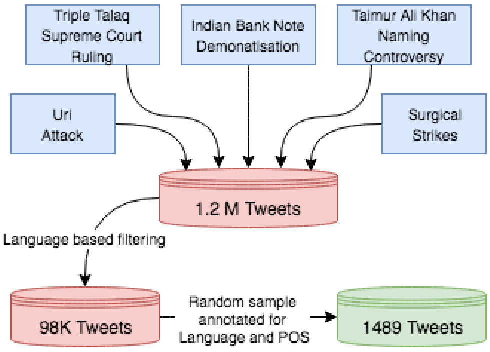
In this work, we present a language and POS-tagged Hindi-English (Hi-En from now on) dataset of 1,489 tweets (33,010 tokens) that closely resembles the topical mode of communication on Twitter. Our dataset is more extensive than any existing code-mixed POS tagged dataset and is rich in Twitter specific tokens such as hashtags and mentions, as well as topical and situational information. We make the entire dataset and our POS tagging model available publicly[2].

2 Related Work

POS tagging is an important stage of an NLP pipeline (Cutting et al., 1992) and has been explored extensively (Toutanova et al., 2003a; Gimpel et al., 2011; Owoputi et al., 2013). However, these models perform poorly on textual content generated on OSNs, including and specially tweets (Ritter et al., 2011). This is due to subtle variations in text generated on OSNs from written and spoken text, such as slack grammatical structure, spelling variations and ad-hoc abbrevi-

[1] Both the terms ”code-mixing” and ”code-switching” are used interchangeably by many researchers

[2] http://precog.iiitd.edu.in/resources.html



Triple Talaq Supreme Court Ruling
Indian Bank Note Demonatisation
Taimur Ali Khan Naming Controversy
Uri Attack
Surgical Strikes
1.2 M Tweets
Language based filtering
98K Tweets
Random sample annotated for Language and POS
1489 Tweets

Figure 1: The dataset creation pipeline.

ations. An elaborate discussion on the differences between tweets and traditional textual content has been done by Ritter et al. (2011).

In addition to the variations between OSN and traditional textual content, code-mixing adds another layer of difficulty (Bali et al., 2014). To bypass these differences, POS taggers have been trained on Hi-En code-mixed posts generated on Facebook (Vyas et al., 2014; Sharma et al., 2016), however, the datasets used for training the models are not available for further experimentation and benchmarking. Only one public dataset of En-Hi code-mixed Twitter posts annotated for POS tags exists (Jamatia and Das, 2016), which comprises of 1,096 tweets (17,311 tokens)[3]. The dataset proposed in this paper is Twitter specific, larger than existing datasets (1,489 tweets, 33,010 tokens) and is event-driven.

3 Dataset Creation

In this section we discuss our data collection methodology and our annotation process. Our data comprises of tweets related to five events, which are (i) the attack by insurgents in the Uri region of Kashmir, India[4], (ii) the Supreme Court ruling that declared Triple Talaq unconstitutional[5], (iii) the Indian banknote demonetization[6], (iv) the Taimur Ali Khan name controversy[7] and (v) the surgical strike carried out by the Indian Army in Pakistan.[8]

3.1 Data Collection and Selection

We first select a set of candidate hashtags related to the five incidents. Using Twitter's streaming API, we collect tweets which contain at least one of these hashtags. For each incident, we collect tweets over a period of 14 days from the day of the incident, collecting 1,210,543 tweets in all.

Previous work has noted that code mixed content forms a fraction of tweets generated, even in multilingual societies (Rudra et al., 2016). To have a high proportion of code mixed tweets in our dataset, we run the language identification model by Sharma et al. (2016) on the tweets and select those which meet all of the following criterion : (i) contains least three Hindi tokens, (ii) contains at least three English tokens, (iii) contains at least 2 contiguous Hindi tokens and (iv) contains at least 2 contiguous English tokens. After this filtering, we are left with 98,867 tweets.

We manually inspect 100 randomly sampled tweets and find that many named entities (NEs) such as 'Kashmir', 'Taimur' and 'Modi' are identified as Hindi. Since manual correction of so many tweets would be difficult, and the problem of misidentifying the language tag of NEs would persist in real life, we include these in our dataset. This misclassification explains the presence of English tweets in our dataset. From the filtered set, we randomly sample 1500 tweets for manual annotation of language and POS. Some of these tweets contain a high number of foreign words (not belonging to English or Hindi). We manually remove such tweets during the annotation process. We maintain the structure of the tweet as it is, and do not split it into multiple sentences. Finally, we tokenize the tweets using twokenizer (Owoputi et al., 2013), which yields 33010 tokens in all.

3.2 Data Annotation

Two bilingual speakers fluent in English and Hindi (one of whom is a linguist) annotate each tweet at the token level for its language and POS. The tags generated by the Language Identifier used for filtering earlier are stripped off. We find that the Language Identifier correctly labeled 82.1% of the tokens, with most misclassification being due to NEs. We note that misclassifications also occur at the boundaries between the Hindi and English part of tweets.

[3] This dataset also comprises of 772 Facebook posts and 762 WhatsApp messages

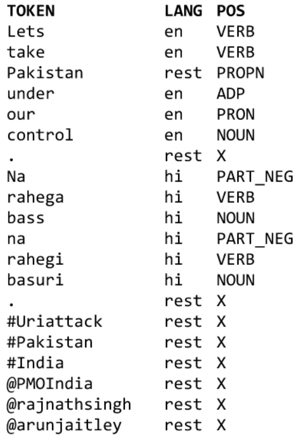
[4] https://reut.rs/2HhBQPg

[5] https://reut.rs/2JDecet

[6] https://reut.rs/2GVKEep

[7] https://bbc.in/2IMPd6Y

[8] https://reut.rs/2EHQZ7g



TOKEN	LANG	POS
Lets	en	VERB
take	en	VERB
Pakistan	rest	PROPN
under	en	ADP
our	en	PRON
control	en	NOUN
.	rest	X
Na	hi	PART_NEG
rahega	hi	VERB
bass	hi	NOUN
na	hi	PART_NEG
rahegi	hi	VERB
basuri	hi	NOUN
.	rest	X
#Uriattack	rest	X
#Pakistan	rest	X
#India	rest	X
@PMOIndia	rest	X
@rajnathsingh	rest	X
@arunjaitley	rest	X

Figure 2: A randomly selected code-mixed tweet from our dataset. The three columns represent the original token, the language tag and the POS tag.

3.2.1 Language Annotation

We use annotation guidelines followed by Bali et al. (2014). Each token is assigned either *hi*, *en* or *rest*. All NEs, Twitter specific tokens (Hashtags and Mentions), acronyms, symbols, words not belonging to Hindi or English, and sub-lexically code-mixed tokens are marked as *rest*. Table 1 and 2 describe the language distribution of our data on a tweet level and token level respectively. Language annotation has a high inter-annotator agreement of 0.97 (Cohen's κ).

3.2.2 Part Of Speech Annotation

Since we look at two different languages, we follow the universal POS set proposed by Petrov et al. (2011) which attempts to cover POS tags across all languages. We reproduce the universal POS set with some alterations, which are (i) We use **PROPN** to annotate proper nouns. We do this to enable further research with this dataset by exploring named entity recognition (NER) which benefits from explicitly labeled proper nouns. All other nouns are tagged as NOUN. (ii) We use **PART_NEG** to annotate Negative Particles. PART_NEG aids in sentiment detection where the presence of a negation word denotes the flipping of sentiment. All other particles are tagged as PART. (iii) We use **PRON_WH** to annotate interrogative pronouns (like where, why, etc.) This shall help in building systems for question detection, another important NLP task. All other pronouns are tagged as PRON.

Language	Tweets
Code-mixed	1077 (72.33 %)
English	343 (23.04 %)
Hindi	69 (4.63 %)
Total	1489

Table 1: Language distribution of tweets. Presence of monolingual tweets is due to errors in the output of the language detection model.

Language	All Tweets	Code-mixed Tweets
English	12589 (38.14 %)	7954 (32.64)
Hindi	9882 (29.94 %)	9093 (37.31)
Rest	10539 (31.93 %)	7323 (30.05)
Total	33010	24370

Table 2: Language distribution of tokens. We observe a fairly balanced spread across the classes.

In the universal set **X** is used to denote foreign words, typos, abbreviations. We also include punctuation under this category. Additionally Twitter-specific tokens hashtags and mentions are also included under X. While (Gimpel et al., 2011) use finer categories for Twitter-specific tokens, we neglect to do so since these tokens can be detected using rule-based features and would artificially boost a POS tagger's accuracy. Figure 2 provides an example of a tweet, and it's corresponding language and POS tag annotation. Inter-annotator agreement for POS tagging was 0.88 (Cohen's κ), all differences were resolved through discussion.

3.3 Data Statistics

Table 3 summarizes the distribution of POS tags in our dataset. We see that there is indeed a high fraction of NEs and that on average, there are 1.84 NEs per tweet. The presence of NEs is confirmed in previous research that event-driven Twitter activity has significant NE content (De Choudhury et al., 2012). We also see a significant amount (421 occurrences) of interrogative pronouns, which in conjunction with 258 occurrences of the '?' symbol signals the presences of inquiries.

4 Experiments

In this section, we demonstrate how our POS-tagged dataset can be used, by building and evaluating an automatic POS tagging model. We present a set of hand-crafted features using which

POS	All tweets	Code Mixed tweets
NOUN	5043 (14.618 %)	3844 (15.773 %)
PROPN	2737 (7.934 %)	1634 (6.705 %)
VERB	5984 (17.346 %)	4566 (18.736 %)
ADJ	1548 (4.487 %)	1116 (4.579 %)
ADV	1021 (2.96 %)	816 (3.348 %)
DET	1141 (3.307 %)	778 (3.192 %)
ADP	2982 (8.644 %)	2229 (9.146 %)
PRON	1456 (4.221 %)	1095 (4.493 %)
PRON_WH	421 (1.22 %)	325 (1.334 %)
PART	1428 (4.139 %)	1122 (4.604 %)
PART_NEG	468 (1.357 %)	399 (1.637 %)
NUM	391 (1.133 %)	309 (1.268 %)
CONJ	809 (2.345 %)	564 (2.314 %)
X	7581 (21.975 %)	5573 (22.868 %)
Total	33010	24370

Table 3: Class wise Part of Speech tag distribution in all Tweets and Code Mixed tweets

POS	POS$_{\text{base}}$	POS$_{\text{base+}}$	POS$_{\text{CRF}}$	POS$_{\text{LSTM}}$
NOUN	72.37	75.95	84.08	72.23
PROPN	81.58	81.68	92.22	80.51
VERB	82.97	79.48	87.84	80.72
ADJ	70.68	69.94	74.92	64.66
ADV	79.26	79.89	82.47	65.92
DET	93.00	95.22	90.50	88.69
ADP	92.92	94.14	93.75	83.75
PRON	87.57	90.91	89.22	83.75
PRON_WH	92.81	93.51	95.60	92.72
PART	78.04	79.93	78.37	73.23
PART_NEG	98.27	98.27	98.27	97.14
NUM	87.32	87.32	90.54	85.51
CONJ	93.55	93.81	93.59	89.23
X	76.11	94.86	98.80	94.51
Total	80.77	85.64	**90.20**	82.51

Table 4: Class wise F$_1$ score (percentage) of different models on the validation set.

our models learn to predict the POS tag of a token. We compare the performance of our models with two naive baselines, POS$_{\text{base}}$ and POS$_{\text{base+}}$. POS$_{\text{base}}$ assigns the most frequent POS tag to a token, as seen in the training data. POS$_{\text{base+}}$ also does the same, but considers the language of the token as well.

For our experiments, we hold out 20% of the data as a validation set. We perform five-fold cross-validation on the remaining 80% for parameter tuning, and report the performance of our models on the validation set in Table 4.

4.1 Model and Features

We attempt to model POS tagging as a sequence labeling task using Conditional Random Field (CRF) and LSTM Recurrent Neural Networks. Previous research has validated the use of CRFs (Toutanova et al., 2003b; Choi et al., 2005; Peng and McCallum, 2006) and LSTM RNNs (Ghosh et al., 2016; Wang et al., 2015) for POS tagging and other sequence labeling NLP tasks.

Our LSTM model has two recurrent layers comprising of 32 bidirectional LSTM cells each. The output of the second layer at each timestep is connected to a softmax layer, used to perform classification over the set of POS tags. Our CRF model is a standard CRF model as proposed by (Lafferty et al., 2001).

We use the following as features for our classifier : **(i)** The current token T, T after stripping all characters which are not in the Roman alphabet (T_{clean}), and converting all characters in T_{clean} to lowercase (T_{norm}) generates three different features, **(ii)** the language tag of T, **(iii)** length of T, **(iv)** Fraction of ASCII characters in T, **(v)** affixes of length 1 to 3, padded with whitespace if needed, **(vi)** a binary feature indicating whether T is titlecased, **(vii)** a binary feature indicating whether T has any upper case character, **(viii)** a binary feature indicating whether there is a non alphabetic character in T and **(ix)** a binary feature indicating whether all characters in T are uppercase.

To prevent overfitting we add a dropout of 0.5 after every layer (for the LSTM model), and L$_1$ and L$_2$ regularization (both models). We perform grid search with 5-fold cross validation to find the optimal values for these parameters.

We supplement the models with a list of rules to detect Twitter specific tokens (such as Hashtags, Mentions, etc.) and Numerals. We follow an approach along the lines of (Ritter et al., 2011) and use regular expressions to make a set of rules for detecting such tokens. Since these are trivial to detect, we omit these tokens while evaluating the performance of the model.

4.2 Results and Error Analysis

Our best model is POS$_{\text{CRF}}$, which achieves an overall F$_1$ score of 90.20% (Table 4). Using the same feature set without language tags led to a slight decrease in F$_1$ score (88.64%). Decrease in POS tagging performance due to language tags is corroborated in previous literature (Vyas et al., 2014). The POS$_{\text{LSTM}}$ model performs poorly (F$_1$ score of 82.51%). We notice that despite using regularization, the model starts overfitting very quickly.

The performance of our POS tagging models across all POS tag categories is shown in Table 4. We find that our POS tagger performs poorest in detecting Hindi adjectives since Hindi has a more relaxed grammatical structure where an adjective may precede as well as follow a noun, e.g.

Tweet: ”U people only talk..no action will be taken! *Aap log darpok ho kewal* Twitter ke sher ho. #UriAttack”

Gloss: ”you people only talk..no action will be taken! *you (aap) people (log) timid(darpok) are(ho)* only(kewal) Twitter of(ke) tiger(sher) are(ho). #UriAttack”

Translation: ”you people only talk..no action will be taken! *you people are timid*, only tiger of Twitter. #UriAttack”

In the above tweet, the adjective ‘timid’ follows the noun ‘people’ instead of the usual format seen in English. A similar trend is observed in adverbs.

5 Discussion

In this data paper, we present a unique dataset curated from Twitter regarding five popular incidents. This dataset differs from previous POS tagged resources both regarding size and lexical structure. We believe that our dataset aids in building effective POS-tagger in order to capture the nuances of Twitter conversation.

We note that our model suffers lower performance for POS tag categories like adjectives and adverbs which follow a different set of grammatical rules for Hindi versus English. In future, we would like to have two POS taggers for differently structured grammar sets and combine them. We also find that our model can detect NEs which is essential when analyzing event-driven tweets. Our dataset therefore also facilitates further research in Named Entity Recognition. We also note the significant amount of interrogative pronouns in our dataset. This suggests that events generate inquiries and questions in the mind of Twitter users.

In future, we would also like to explore building other downstream NLP tools such as Parsers or Sentiment Analyzers which make use of POS tags using our dataset and refined versions of our POS tagging model.

References

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. “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*. Association for Computational Linguistics, pages 116–126. https://doi.org/10.3115/v1/W14-3914.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 355–362.

Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, pages 133–140.

Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, pages 241–244.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291* .

Kevin Gimpel, Nathan Schneider, Brendan O’Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 42–47.

Anupam Jamatia and Amitava Das. 2016. Task report: Tool contest on pos tagging for code-mixed indian social media (facebook, twitter, and whatsapp) text @ icon 2016. In *Proceedings of ICON 2016*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML ’01, pages 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using conditional random fields. *Information processing & management* 42(4):963–979.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086* .

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1524–1534.

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

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Shrivastava, Radhika Mamidi, and Dipti M Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. In *Proceedings of NAACL-HLT*. pages 1340–1345.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003a. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL '03, pages 173–180. https://doi.org/10.3115/1073445.1073478.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003b. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 173–180.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *EMNLP*. volume 14, pages 974–979.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. *arXiv preprint arXiv:1511.00215* .