SyMCoM - Syntactic Measure of Code Mixing
A Study Of English-Hindi Code-Mixing

Prashant Kodali[†] **Anmol Goel**[†] **Monojit Choudhury**[‡]
Manish Shrivastava[†] **Ponnurangam Kumaraguru**[†]
[†]International Institute of Information Technology Hyderabad
[‡]Microsoft Research, India
{prashant.kodali, anmol.goel}@research.iiit.ac.in
monojitc@microsoft.com, {m.shrivastava, pk.guru}@iiit.ac.in

Abstract

Code mixing is the linguistic phenomenon where bilingual speakers tend to switch between two or more languages in conversations. Recent work on code-mixing in computational settings has leveraged social media code mixed texts to train NLP models. For capturing the variety of code mixing in, and across corpus, Language ID (LID) tags based measures (CMI) have been proposed. Syntactical variety/patterns of code-mixing and their relationship vis-a-vis computational model's performance is under explored. In this work, we investigate a collection of English(en)-Hindi(hi) code-mixed datasets from a syntactic lens to propose, $SyMCoM$, an indicator of syntactic variety in code-mixed text, with intuitive theoretical bounds. We train SoTA en-hi PoS tagger, accuracy of 93.4%, to reliably compute PoS tags on a corpus, and demonstrate the utility of $SyMCoM$ by applying it on various syntactical categories on a collection of datasets, and compare datasets using the measure.

1 Introduction

Code-mixing refers to mixing of linguistic units and structures from multiple languages in a single utterance and/or conversation (Myers-Scotton, 1997). The complexity of code-mixing can be intuitively understood as the degree of structural interleaving between the languages at the level of the lexicon and morpho-syntax (Myers-Scotton, 1997), and also at the level of pragmatic and sociolinguistic functions of code-mixing in a linguistic community (Begum et al., 2016; Annamalai, 2001; Malhotra, 1980). It is an important notion that is linguistically well-studied and provides insights into cognitive and cultural aspects of human language. Additionally, quantification of this complexity has recently attracted attention of computational linguists because studies have shown that the performance of the same model can widely vary on different code-mixed corpora. As a result, different metrics of complexity of code-mixing have been proposed such as CMI, Ratio-based measures time-course measures and memory-based measures (Guzmán et al., 2017; Gambäck and Das, 2016). But as Srivastava and Singh (2021) points out, these metrics are limited, partly because they are primarily based on language switch patterns at the token level, being completely agnostic to structural features.

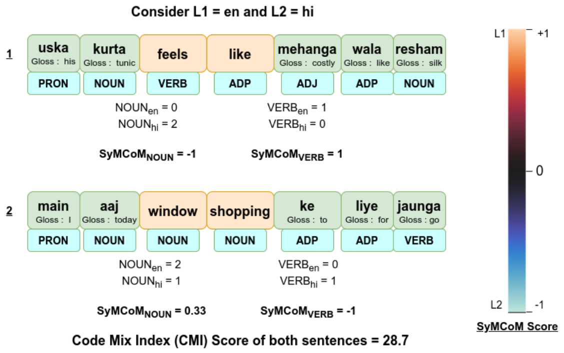


Figure 1: Two sentences, having same language patterns, but the syntactic nature of the switched units are different - VERB is switched in Ex 1, NOUNS are switched in Ex 2.

For instance, Figure 1 shows $en-hi$ code-mixed sentences with the same language tag distribution, but in Example (1), verb is switched, while in (2) nouns are switched. The former seems much more complex and difficult to process cognitively and computationally, than the latter, where switching seems to be an extension of noun-borrowing (Bali et al., 2014). Motivated by such cases, in this paper, we ask the following research question: *Can syntactic category information be deduced and used as a measure of structural complexity of a code-mixed sentence and corpus?* We attempt to tackle this question by formulating: **S**yntactic **M**easure of **Co**de **M**ixing $SyMCoM$, a simple metric that encodes the distributional difference of various syntactic categories across languages in a sentence. $SyMCoM$ can be computed for any corpora, as long as there is a reasonably accurate POS tagger

Findings of the Association for Computational Linguistics: ACL 2022, pages 472 - 480
May 22-27, 2022 ©2022 Association for Computational Linguistics

for the code-mixed language pair.

Through empirical studies of several existing en-hi code-mixed corpora we provide, for the first time, a strong quantitative evidence in support of a widely held theoretical notion that Open class categories (e.g., noun, adjectives) are more likely to be switched than the closed class categories (e.g., pronouns, verbs) within a sentence. Further, we show that different corpora have significantly different distribution of $SyMCoM$ values.

2 SyMCoM: Syntactic Measure of Code Mixing

A quantitative measure of syntactic variation in code mixing patterns should ideally encode: **a) Category of Switch** i.e whether or not a PoS tag or syntactic category is switched?; **b) Degree / Contrast** : If a syntactic unit is switched, what is the level of contrast between $L1$ and $L2$ for that unit?

To encode the aforementioned properties, we propose a **S**yntactic **M**easure of **Co**de **M**ixing ($SyMCoM_{SU}$), which is defined as:

$$SyMCoM_{SU} = \frac{(Count_{SU_{L1}}) - (Count_{SU_{L2}})}{\sum_{i=1}^{2} Count_{SU_{Li}}} \qquad (1)$$

Here, SU is a syntactic unit; for this study, we will assume that SU represent word-level syntactic categories namely Parts-of-Speech (POS) tags such as Nouns and Verbs, or a class of PoS tags such as Open and Closed classes. $Count_{SU_{Li}}$ represents the count of the syntactic unit SU for language L_i $(i \in \{1, 2\})$ within a sequence of words code-mixed between languages L_1 and L_2. Without loss of generality, we will consider this sequence to be a sentence, though it could be an utterance, paragraph or even a document. $SyMCoM_{SU}$ score is bounded between [-1,1] and defined only for SUs that occur at least once in the sentence.

The polarity of $SyMCoM_{SU}$ indicates the language, among $L1$ and $L2$, that is contributing higher number of tokens for a particular SU, and its absolute value captures the degree of skew towards a particular language. If $SyMCoM_{SU}$ is closer to zero, it indicates that the contribution of $L1$ and $L2$ for SU is balanced.

We define the $SyMCoM_{sent}$ score for a sentence, as the weighted average of absolute $SyMCoM_{SU}$ scores for all SU, where the weights are the fraction of tokens in the sentence belonging to an SU.

$$SyMCoM_{sent} = \sum_{SU} \frac{Count_{SU}}{len} \times |SyMCoM_{SU}| \qquad (2)$$

$SyMCom_{sent}$ is bounded between [0,1]. Values closer to zero indicate that L_1 and L_2 contribute nearly equally for most types of SUs in the sentence, whereas values close to 1 indicate that in the sentence each SU is majorly contributed by a single language. Note that while a low $SyMCoM_{sent}$ implies that the tokens in a sentence are nearly equally contributed by the two languages, a high $SyMCoM_{sent}$ does not say anything about the language distribution of the tokens.

$SyMCoM_{sent}$ can be averaged over the corpus to capture the syntactic variation at a corpus level:

$$SyMCoM_{corpus} = \sum_{sent} \frac{SyMCoM_{sent}}{\#\ sentences\ in\ corpus} \qquad (3)$$

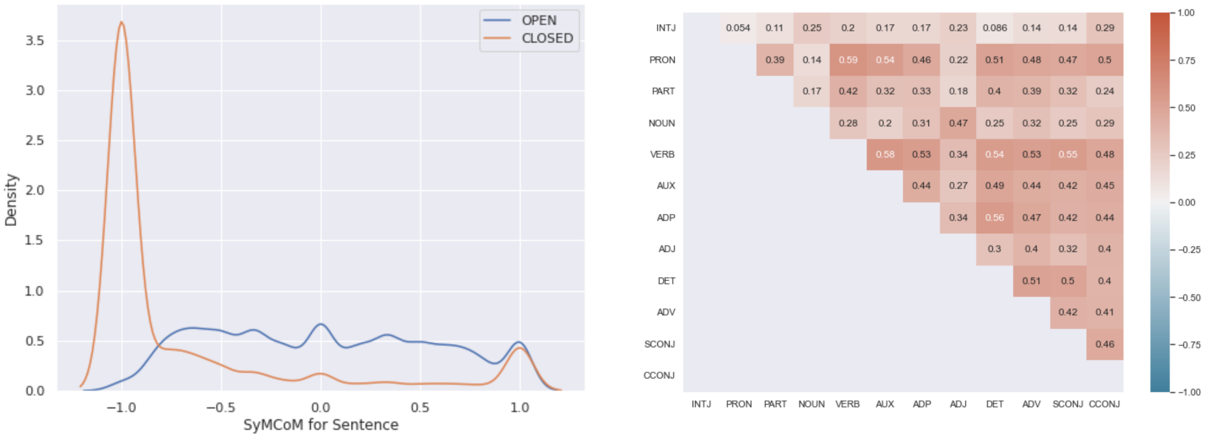
Equation 1 can be extended to any arbitrary subset of POS categories, of which the Open Class (content words - Noun, Adjectives, Verbs, etc.) and Closed Class (function words - Adpositions, Pronouns, Demonstratives, etc.) are of special interest; we will refer to these as $SyMCoM_{OPEN}$ and $SyMCoM_{CLOSED}$ respectively.

In Figure 1, $SyMCoM_{SU}$ scores are computed for two en-hi code-mixed sentences, whose CMI scores are equal. For each utterance, we calculate the number of nouns and verbs belonging to L_1 = en and L_2 = hi. In Example 1, $SyMCoM_{NOUN}$ = -1, indicating that L_2 is contributing all the Nouns in this sentence. The opposite polarity of $SyMCoM_{VERB}$ indicates that all the verbs are contributed by L_1.

3 Experiments & Discussion

To demonstrate the utility of the proposed $SyMCoM$ measure, we analyse a) en-hi code-mixed corpus; b) compare $SyMCoM_{sent}$ $SyMCoM_{corpus}$ score across different datasets. To compute $SyMCoM$ scores we need token wise LID and PoS tags. We use pre-trained character level BiLSTM Language ID tagger released by (Bhat et al., 2018) for obtaining token wise LIDs. We train our PoS tagger using the en-hi Universal Dependency dataset released by (Bhat et al., 2017, 2018), which used Universal Dependency tagset (de Marneffe et al., 2014).

Use of pre-trained BiLSTM might not be good for few use cases.



a) SyMCoM for OPEN and CLOSED Categories

b) Correlation of SyMCoM scores for PoS Tags

Figure 2: (a) The peak of the curves indicates that CLOSED class words are commonly used in a single language while OPEN class words are spread out, hence, are contributed by both languages. (b) $SyMCoM_{SU}$ score of VERB (Open Class) is highly correlated with Closed class tags.

3.1 en-hi Code Mix PoS tagger

The GLUECoS (Khanuja et al., 2020) and LinCE (Aguilar et al., 2020) benchmarks indicate that multilingual transformer based encoder models - mBERT (Devlin et al., 2019) have matched or outperformed the SoTA on the specific PoS tagging tasks, while showing sub par performance on more complex semantic tasks such as Sentiment Analysis and Natural Language Inference.

Model	Accuracy
Bhat et al. (2018)	91.9%
Mod. mBERT (Khanuja et al., 2020)	88.06%
XLM-R	**92.75%**

Table 1: PoS Tagger Performance

We fine-tune XLM-R (Conneau et al., 2020) to obtain best-performing PoS Tagger. In Table 1 we compare accuracy of our best-performing fine-tuned XLM-R model against previous results reported in GLUECoS benchmark, and Bhat et al. (2018). In addition to accuracy, we also analyse the class wise performance, and we note that ADV, INTJ, PROPN have f1 lower than 0.85. $SyMCoM$ measure depends on the accuracy of the PoS tags and the potency of PoS tagger impacts the usability of the score. We recommend that for $SyMCoM$ scores, the corresponding accuracy of detecting syntactic unit shall be taken into account, and $SyMCoM$ scores be computed for syntactic units which can be detected with a higher accuracy. Further training details for the PoS tagger are listed in Appendix A.

Using the LID and PoS tags, for each Syntactic Unit considered, the language specific counts are computed - SU_{L_i}. $SyMCoM_{SU}$ scores for particular syntactic unit are then calculated using the counts SU_{L_i} as mentioned in Equation 1.

3.2 Analysis of en-hi Code Mixed Corpus

To demonstrate the utility of the proposed $SyMCoM$ measure, we analyse en-hi code-mixed corpus. We collect publicly available code-mixed en-hi datasets released for various tasks: shared tasks, code mixed benchmarks (GLUECoS, LINCE), text classification, Machine translation, among others- remove any monolingual sentences, and created a corpus of 55,474 sentences, details of the datasets used are in Appendix C. $SyMCoM$ scores, along with CMI score (Gambäck and Das, 2016), are computed for the collected corpus, and compared.

Figure 2(a) shows the distribution of $SyMCoM_{SU}$ scores for Open and Closed class units. The skewed nature of $SyMCOM_{CLOSED}$ indicates that Closed class words are not mixed, and are provided by either $L1$ or $L2$. $SyMCoM_{OPEN}$, on the other hand, is more spread out indicating that in code-mixed sentences the Open class tokens are contributed by both *en* and *hi*. Figure 2(b) indicates correlations of $SyMCoM$ scores for all the PoS categories. Higher correlation scores indicate the tendency of the particular PoS tag pair to switch together. Similar to Figure 2(a), the correlations indicate that closed class tokens are from the same language.

2 major observations

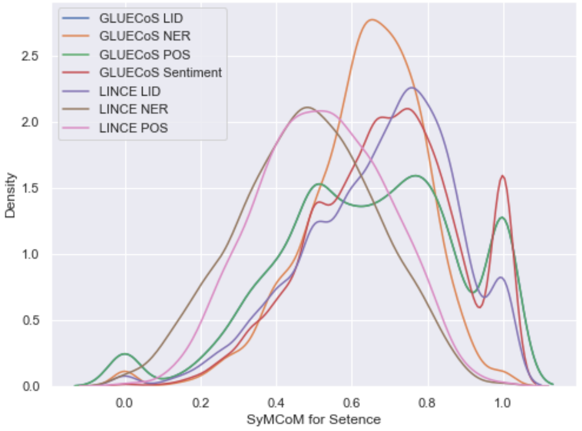


Figure 3: $SyMCoM_{sent}$ scores for various benchmark datasets. The plot represents the syntactic variation across benchmark datasets which encode the switching within PoS tag categories.

Interestingly, verb is also highly correlated with closed class categories. The high correlation can be attributed to the fact that the finite verbs, along with closed class words, govern syntactical structure of a sentence, similar to the notion of matrix language. According to (Joshi, 1982), certain categories including pronouns, adpositions and finite verbs cannot be switched *from* the matrix language. What's a matrix language?

Figure 3 shows the $SyMCoM_{sent}$ distribution over sentences for several code-mixed corpora taken from the GLUECoS (Khanuja et al., 2020) and LINCE (Aguilar et al., 2020) benchmarks. Clearly, the 7 corpora has distinct $SyMCoM$ signatures. While LINCE POS and NER has very similar normal-like distributions with mean 0.5, all the other datasets seem to be right skewed showing less syntactic complexity. Most GLUECoS datasets show a bimodal distribution with a major peak between 0.6 and 0.8, and a minor peak at 1, indicating that a significant fraction of the sentences have syntactically simple code-mixing patterns, and most being only moderately syntactically complex. GLUECoS POS dataset though have four peaks including one at 0 implying a more complex and diverse set of sentences.

ELI5 this >>>>>>

Table 2 reports the $SyMCoM_{corpus}$ and CMI score. Datasets with seemingly similar CMI scores (LINCE LID and LINCE NER), have different $SyMCoM_{corpus}$ scores, indicating that $SyMCoM$ is capturing syntactic property of datasets not captured in CMI scores. Figure 4 shows the $SyMCoM_{SU}$ scores for each PoS tag, and compared for the benchmark datasets. We average the $SyMCoM_{SU}$ scores for each PoS tag over the dataset, by averaging the absolute value of $SyMCoM_{SU}$ score. Nouns and verbs are mixed the most, across all datasets while other PoS tags remain largely monolingual.

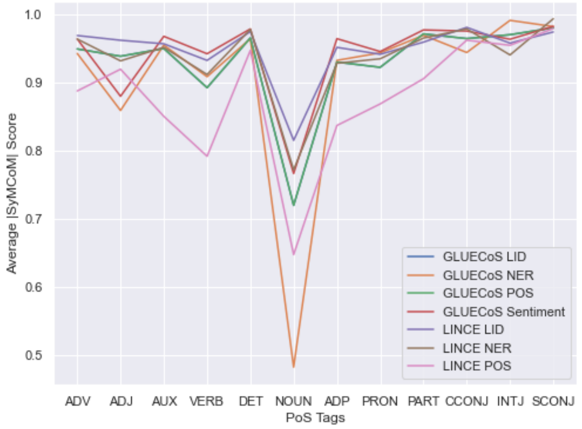


Figure 4: $SyMCoM_{sent}$ scores for various benchmark datasets for individual PoS tags. The plot represents mixing specific to each PoS tag. Across benchmarks, NOUN is highly switched, followed by VERB. But other PoS tags are largely monolingual.

Dataset	$SyMCoM_{corpus}$	CMI
LINCE LID	0.67	22.68
GLUECoS LID	0.64	78.26
LINCE POS	0.52	28.04
GLUECoS POS	0.64	68
LINCE NER	0.48	25.26
GLUECoS NER	0.63	133
GLUECoS Sentiment	0.69	72.8

Table 2: $SyMCoM_{corpus}$ and CMI measures for benchmark En-Hi datasets. $SyMCoM_{corpus}$ is bounded between [0,1] while CMI > 0. For datasets with similar CMI scores, $SyMCoM_{corpus}$ is able to distinguish datasets.

4 Conclusion & Limitations

In this work, we have proposed $SyMCoM$, a syntax-aware measure of code-mixing, to analyze code-mixed corpora from a syntactic perspective. Our analysis confirms a few important tenets of the matrix language theory, including the fact that CLOSED class categories and (finite) verbs are less likely to be switched. Additionally, we have trained a English-Hindi (Hinglish) PoS Tagger using XLM-R which is able to achieve state-of-the-art-results. <<< IMPORTANT

$SyMCoM$ relies on the strength of PoS tagger and LID tagger . The errors made by the tagger would propagate into the subsequent analysis thus adding noise to the $SyMCoM$ scores as well. Extending $SyMCoM$ to code-mixing between 3 or more languages and to deeper syntactic structures (nested phrases) are left as part of future work.

<<< can make potential improvement here.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

E. Annamalai. 2001. Managing multilingualism in india: Political and linguistic manifestations.

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*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

Somnath Banerjee, M. Choudhury, K. Chakma, S. Naskar, Amitava Das, Sivaji Bandyopadhyay, and P. Rosso. 2020. Msir@fire: A comprehensive report from 2013 to 2016. *SN Comput. Sci.*, 1:55.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M. Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1644–1650, Portorož, Slovenia. European Language Resources Association (ELRA).

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2017. Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 324–330, Valencia, Spain. Association for Computational Linguistics.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. 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 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

K. Chakma and Amitava Das. 2016. Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets. *Computación y Sistemas*, 20:425–434.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Amitava Das. Tool contest on pos tagging for code-mixed indian social media (facebook, twitter, and whatsapp) text @ icon 2016.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for Modeling Code-Switching Across Corpora. In *Proc. Interspeech 2017*, pages 67–71.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor detection in english-hindi code-mixed social media content : Corpus and baseline system.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

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*, pages 3575–3585, Online. Association for Computational Linguistics.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sunita Malhotra. 1980. Hindi-english, code switching and language choice in urban, uppermiddle-class indian families.

C. Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task @icon-2017. *CoRR*, abs/1803.06745.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text.

Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019. Stance detection in code-mixed Hindi-English social media data using multi-task learning. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–5, Minneapolis, USA. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14, Online. Association for Computational Linguistics.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018a. A corpus of english-hindi code-mixed tweets for sarcasm detection.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018b. An english-hindi code-mixed corpus: Stance annotation and baseline system.

Appendix

A PoS Tagger Training and Performance Details

We conducted preliminary experiment for PoS tagging using - XLM-R, mBERT, IndicBERT (Kakwani et al., 2020), MURiL (Khanuja et al., 2021), and results indicate that XLM-R with normalised inputs (romanised Hindi tokens are converted to their Devanagari counterpart), outperforms other models. We run further fine-tuning experiments on XLM-R to obtain optimal performing PoS Tagger. We tested three approaches : **Method 1**: Leverage Transfer from larger Monolingual UD datasets by fine tuneing XLM-R on Hindi and English Monolingual UD datasets PoS tagging , followed by fine tuning on en-hi UD dataset; **Method 2**: Directly fine tune XLM-R on UD Code mix hi-en dataset, using the un-normalised tokens i.e romanised hindi tokens are in roman script; **Method 3**: Directly fine tune XLM-R on UD Code mix hi-en dataset, using the normalised tokens i.e romanised hindi tokens are in Devanagari script

Split	Num. of Samples	Num of Tokens
Train	1,448	20,203
Dev	225	3,411
Test	225	3,295

Table 3: Statistics of en-hi UD Dataset

Table 3 shows that the size of dataset used for training and validation isn't large, hence, for the best performing model, we try to assess the variation in results due to different seeds and data shuffling, show in Figure 5. Highest accuracy achieved by the model is 93.34%, with $\mu = 92.75\%$ and $\sigma = 0.35$.

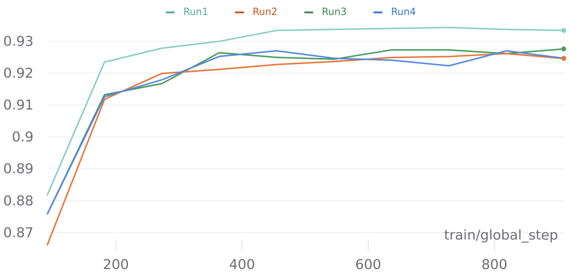
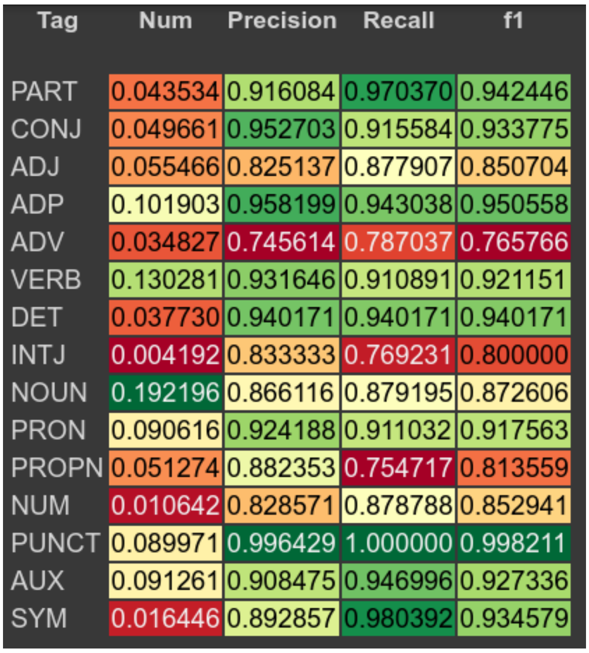


Figure 5: Variation in PoS tagging performance for different values of random seeds and data shuffling. Best accuracy is 93.4%,least being 92.4%.

All the results are reported on the normalised inputs. Fine-tuned XLM-R model outperforms the previous results reported on this dataset. Trained model checkpoint can be found at `https://huggingface.co/prakod/en-hi-pos-tagger-symcom`.



Tag	Num	Precision	Recall	f1
PART	0.043534	0.916084	0.970370	0.942446
CONJ	0.049661	0.952703	0.915584	0.933775
ADJ	0.055466	0.825137	0.877907	0.850704
ADP	0.101903	0.958199	0.943038	0.950558
ADV	0.034827	0.745614	0.787037	0.765766
VERB	0.130281	0.931646	0.910891	0.921151
DET	0.037730	0.940171	0.940171	0.940171
INTJ	0.004192	0.833333	0.769231	0.800000
NOUN	0.192196	0.866116	0.879195	0.872606
PRON	0.090616	0.924188	0.911032	0.917563
PROPN	0.051274	0.882353	0.754717	0.813559
NUM	0.010642	0.828571	0.878788	0.852941
PUNCT	0.089971	0.996429	1.000000	0.998211
AUX	0.091261	0.908475	0.946996	0.927336
SYM	0.016446	0.892857	0.980392	0.934579

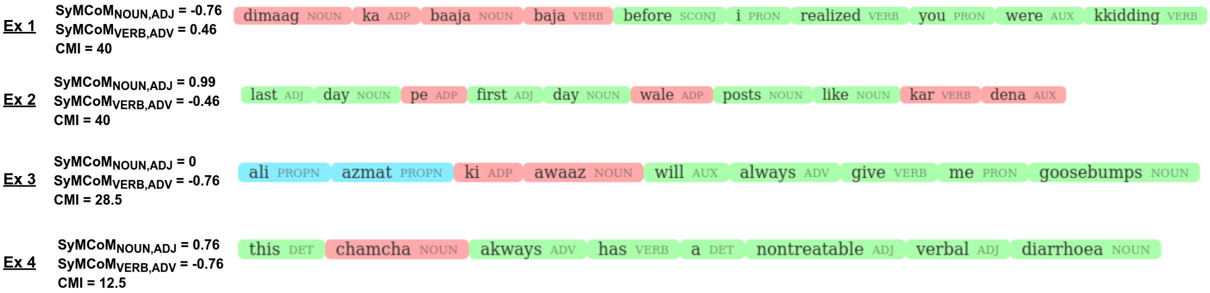
Figure 6: Class wise performance of PoS tagger. We can see that the certain classes like ADV, INTJ, PROPN have lowwer performance compared to other classes.

B Example Sentences from the collection of Datasets

In Figure 7 examples are selected from the corpus on which the $SyMCoM$ scores were computed using the LID and POS tagger outputs. We contrast the CMI score against $SyMCoM$ scores using these examples. Example (1) and (2) have same CMI score, however syntactic signature of code-mixing is quite distinct. In example (2), nouns and adjective are contributed by en, while in example (1) nouns are contributed by hi. Similarly in example (3) $SyMCoM$ score for [NOUN,ADJ] is zero indicating that both en and hi contribute equal number of tokens belonging to the syntactical category of [NOUN,ADJ].

C Dataset Sources

In Table 4, we list all the sources used to construct our 55K sentence corpus of English-Hindi code-mixing. The data is representative of a wide variety of code-mixing including Hate Speech, Stance



Ex 1 SyMCoM$_{NOUN,ADJ}$ = -0.76 SyMCoM$_{VERB,ADV}$ = 0.46 CMI = 40

dimaag NOUN ka ADP baaja NOUN baja VERB before SCONJ i PRON realized VERB you PRON were AUX kkidding VERB

Ex 2 SyMCoM$_{NOUN,ADJ}$ = 0.99 SyMCoM$_{VERB,ADV}$ = -0.46 CMI = 40

last ADJ day NOUN pe ADP first ADJ day NOUN wale ADP posts NOUN like NOUN kar VERB dena AUX

Ex 3 SyMCoM$_{NOUN,ADJ}$ = 0 SyMCoM$_{VERB,ADV}$ = -0.76 CMI = 28.5

ali PROPN azmat PROPN ki ADP awaaz NOUN will AUX always ADV give VERB me PRON goosebumps NOUN

Ex 4 SyMCoM$_{NOUN,ADJ}$ = 0.76 SyMCoM$_{VERB,ADV}$ = -0.76 CMI = 12.5

this DET chamcha NOUN akways ADV has VERB a DET nontreatable ADJ verbal ADJ diarrhoea NOUN

Figure 7: Example sentences demonstrate that the sentences having different SyMCoM scores, and SyMCoM scores sentences can distinguish between sentences with similar CMI. Color indicate LID tags, where in en , hi , ne

Detection, Humor Detection and conversational systems.

LINCE Benchmark	(Aguilar et al., 2020)
GLUECoS Benchmark	(Khanuja et al., 2020)
Sentiment Analysis	(Prabhu et al., 2016)
Semeval-2020 Sentiment Analysis	(Patwa et al., 2020)
Machine Translation	(Dhar et al., 2018)
Aggression Detection Shared Task	(Kumar et al., 2018)
Hate Speech Detection	(Bohra et al., 2018)
Stance Detection	(Swami et al., 2018b)
Stance Detection	(Sane et al., 2019)
Sarcasm Detection	(Swami et al., 2018a)
Humor Detection	(Khandelwal et al., 2018)
Code Mixed Goal Oriented Conversation Systems	(Banerjee et al., 2018)
ICON 2015-2016 PoS LID Contest	(Das)
FIRE 2013-16 Tasks	(Banerjee et al., 2020)
Information Retrieval	(Chakma and Das, 2016)
Sentiment Analysis	(Patra et al., 2018)

Table 4: English-Hindi Code mix datasets used to construct the 55K sentence corpus. $SyMCoM$ scores are calculated on the collected sentences. LINCE and GLUECoS benchmark datasets are used to contrast syntactic variety of code mixing across datasets.