BITS Pilani at HinglishEval: Quality Evaluation for Code-Mixed Hinglish Text Using Transformers

Shaz Furniturewala, Vijay Kumari, Amulya Ratna Dash, Hriday Kedia, Yashvardhan Sharma

BITS Pilani,Pilani,Rajasthan

`(f20200025,p20190065)@pilani.bits-pilani.ac.in`
`(p20200105,f20190964,yash)@pilani.bits-pilani.ac.in`

Abstract

Code-Mixed text data consists of sentences having words or phrases from more than one language. Most multi-lingual communities worldwide communicate using multiple languages, with English usually one of them. Hinglish is a Code-Mixed text composed of Hindi and English but written in Roman script. This paper aims to determine the factors influencing the quality of Code-Mixed text data generated by the system. For the HinglishEval task, the proposed model uses multi-lingual BERT to find the similarity between synthetically generated and human-generated sentences to predict the quality of synthetically generated Hinglish sentences.

1 Introduction

The term "Code-Mixing" refers to mixing words or phrases from different languages into a single text or speech utterance. It embeds linguistic units from one language, such as phrases, words, and morphemes, into an utterance from another language. An example of the Code-Mixed data can be seen in the Figure 1 (Srivastava and Singh, 2021b). In countries where bilingualism is a common practice, we often see people naturally switching between the two languages. A significant challenge to research is that there are no formal sources like books or news articles in Code-Mixed languages, and studies have to rely on sources like Twitter or messaging platforms. Generating and evaluating the available or produced data without a baseline is primarily reliant on people who are fluent in both languages.

Furthermore, present language models are ineffective in Code-Mixed situations, where morphemes, words, and phrases from one language are embedded in the other. As Code-Mixing has long been a way of communication in a multi-cultural,

multi-lingual society, the next generation of AI bots should be able to understand Code-Mixed text.

The inherent challenges with the code-mixed data make the widely popular evaluation metrics like BLEU (Bilingual Evaluation Understudy Score) and WER (Word Error Rate) less effective. With the given task, the main objective is to propose and develop new strategies that address the overall need for quality evaluation of the generated Code-Mixed text.

This paper aims to assess the quality of the generated Hinglish text. The proposed model uses the transformer-based multi-lingual BERT (Pires et al., 2019) to obtain the embeddings of the Hindi, English, and Hinglish text. A similarity score is computed between synthetically generated Hindi and English text and human-generated Hindi and English sentences. This score will serve as the sentence's agreement or disagreement factor.
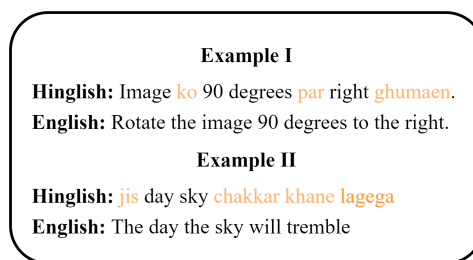

Example I

Hinglish: Image ko 90 degrees par right ghumaen.
English: Rotate the image 90 degrees to the right.

Example II

Hinglish: jis day sky chakkar khane lagega
English: The day the sky will tremble

Figure 1: Examples of Hinglish and English sentences

2 Related Work

Developing methodologies and resources for diverse Natural Language Processing applications incorporating multi-lingual and Code-Mixed languages has recently become popular. Few of them are word-embedding (Chen and Cardie, 2018), question answering system (Raghavi et al., 2015; Kumari et al., 2022), Code-Mixed text generation

(Pratapa et al., 2018) and Code-Mixed language modeling (Winata et al., 2018).

Various techniques were used to generate the Code-Mixed data, including matrix frame language theory, semi-supervised approach (Gupta et al., 2020), using dependency parsing (Jain et al., 2021), equivalent constraint theory (Pratapa et al., 2018) and Generative Adversarial Networks (Gao et al., 2019). A Metric Independent Evaluation Pipeline (MIPE) (Garg et al., 2021) considerably enhances the correlation among evaluation measures and human assessments of the generated Code-Mixed data. In the candidate's Hinglish phrase, MIPE minimises spelling differences and language switches for evaluation. Based on the significance of the terms missing from the candidate Hinglish sentence, deduct credit from the evaluation score. By arranging the candidates as well as the reference sentences into the phrases and using the paraphrasing ability, it also deals with the issue of having a restricted amount of reference sentences.

3 Dataset

HinGE(Srivastava and Singh, 2021a), a recently proposed dataset, is used for the HinglishEval task. The dataset contains 1976 English-Hindi sentence pairs and corresponding synthetic and human-generated Hinglish sentences.

Human Generated Sentences: Each English-Hindi sentence pair has at least two Hinglish sentences, with a total of 6694 such sentences.

Synthetically Generated Sentences: Using two different algorithms(WAC and PAC), synthetic Hinglish sentences are generated for each English-Hindi sentence pair. Each sentence is then given a quality rating by two human annotators. There are 2766 such sentences. For each sentence, the rounded off average of the two quality ratings is provided under the label of 'Average Rating' and the absolute difference of their scores is provided as 'Disagreement'. An instance of the HinGE dataset (Srivastava and Singh, 2021a) can be seen in Table 1.

4 Proposed Model

The proposed approach used a two-step procedure for both the rating and disagreement prediction tasks. The first step is to fine-tune multi-lingual BERT, a language model that has been pre-trained on 104 languages and is used to classify and evaluate disagreements further on. The second step used
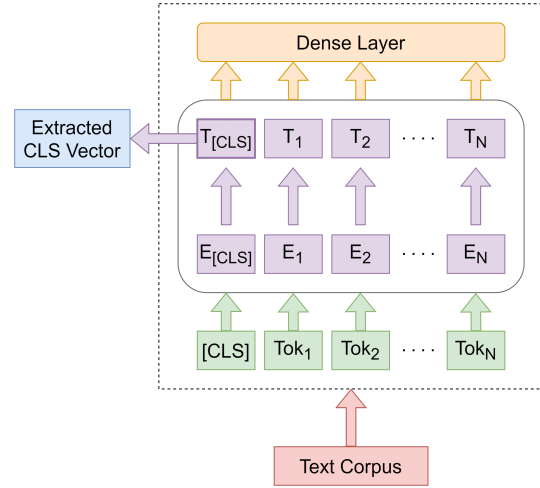

Figure 2: Extraction of CLS Vectors

the deep semantic features obtained from multi-lingual BERT for various phrase categories to train a classifier neural network.

4.1 Multi-lingual BERT

We employed the BERT-base-multi-lingual-cased model, a modified pre-trained BERT encoder that is pretrained in a self-supervised manner using the largest Wikipedias with the goal of masked language modeling. Multi-lingual BERT allows us to provide it with two sentences as input in the form:

[CLS] Sentence A [SEP] Sentence B [SEP]

Here, the CLS and SEP tokens are special tokens that allow BERT to recognize the beginning of an input and a separation between two different input sections. The central SEP token ensures that BERT knows there are two different sentences in the input.

Based on the dataset, we had four models for different types of sentences—English, Hindi, Synthetic Hinglish, and Human Hinglish. We ran the language model four times for each sub-tasks with various sorts of sentence pairs each time. The first two models trained had Synthetic Hinglish sentences as Sentence B, with Sentence A being the corresponding English sentence for model 1 and Hindi sentence for model 2. The next two models did the same, with Human Hinglish sentences paired with English sentences for model 3 and Hindi sentences for model 4. This process is repeated twice for rating classification and again for disagreement classification. The proposed model employs the BERT AutoTokenizer to tokenize the inputs and the Adam optimizer to train at a learning

English	Hindi	Human-Generated Hinglish	WAC	PAC
The reward of goodness shall be nothing but goodness	अच्छाई का बदला अच्छाई कए अलावा ओर कया हो सकता ह?	The reward of achai shall be nothing but achai.	reward ka badla reward ke nothing aur kya ho sakta hai	reward of goodness goodness ke siva aur kya ho sakta hai
		Goodness ka badla goodness ke siva aur kya ho sakta hai.	**Rating1: 7** **Rating2: 4**	**Rating1: 9** **Rating2: 7**
		Achai ka badla shall be nothing but achai.		

Table 1: An Instance of the HinGE dataset

rate of 1e-6 for five epochs.

We extracted the deep semantic text features from each model using the BERT source code. The CLS token's feature vector is extracted from the output results of the last hidden layer, which is a 768-dimensional deep semantic feature of the legal language. We chose the CLS token, also known as the Classification token, as it has the fixed embedding that appears at the beginning of every sentence. Since all words infer the output of this token in the phrase, this CLS vector provides BERT's understanding of the sentence, which is particularly beneficial for a sentence classification task. The extraction process can be seen in Figure 2, where $T_{[CLS]}$ is the CLS vector output of the last hidden layer of multi-lingual BERT extracted from the model.

4.2 Classifier Neural Network

We obtained four sets of feature vectors from multilingual BERT for each sub-task. Two of them had dimensions of (2766, 768) and used synthetic Hinglish words as input. The other two used human-generated sentences and had dimensions of (6694, 768). To reduce the dimensionality of the latter, we averaged the vectors corresponding to each English-Hindi sentence pair resulting in a set of dimensions (1976, 768).

For each synthetic sentence, we concatenated the first two sets of vectors, and corresponding to each vector in this set, we appended its respective human vector. After combining the four-vector sets, we had an input of size (2766, 3072) for each sub-task.

The proposed model is trained with two fully connected neural networks, one for each sub-task, using these concatenated vector sets as their respective inputs. Both neural networks had two Linear hidden layers of dimensions (3072, 1536) and (1536, 768) with a final layer of size (768, 10). After each layer, a Rectified Linear activation function is being used. Binary Cross-Entropy loss is the chosen loss function. Adam optimizer were used along with a learning rate of 5e-6. Ten training epochs were used to train the disagreement classifier compared to just three for the ratings classifier. The entire procedure is illustrated in Figure 3, where the four concatenated CLS vectors are passed as input to a classifier neural network.
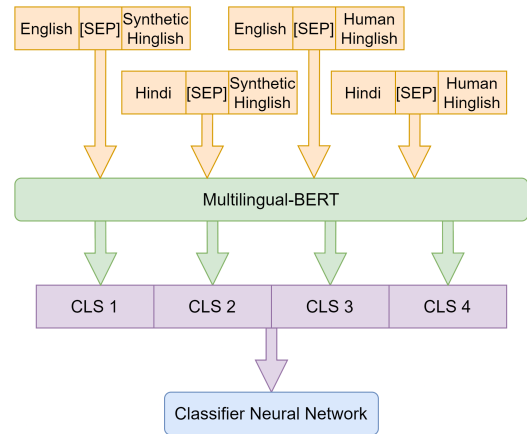

Figure 3: Proposed Model Architecture

	Model	Sub-Task 1(Average Rating)			Sub-Task 2(Disagreement)	
		F1	CK	MSE	F1	MSE
Validation	Baseline [1]	0.09504	-0.01530	15.000	0.15541	17.000
	Proposed Model	0.23493	0.06515	3.000	0.19400	4.000
Test	Baseline [1]	0.26637	0.09922	2.000	0.14323	5.000
	Proposed Model	0.21796	0.07337	3.000	0.24252	4.000

Table 2: Validation and Test results on the HinGE dataset

5 Results and Evaluation

5.1 Results

The evaluation metrics used are F1 Score, Cohen's Kappa (CK), and Mean Square Error (MSE). The submissions created by the model achieved rank 4 in rating prediction and rank 2 in disagreement prediction, not accounting for baseline scores, based on F1 Score, Cohen's Kappa, and mean squared error. The proposed approach results can be seen in Table 2, compared with the baseline scores.

We attained F1 scores of 0.218 on Rating classification and 0.242 on Disagreement classification after training over 2766 synthetic sentences and testing over 791 synthetic sentences.

6 Conclusion and Future Work

This paper used a two-step approach to solve the text classification problem. For each pair of phrase types, deep semantic text features were initially extracted using multilingual BERT as CLS vectors. These vectors were then properly combined, and processed by a fully-connected classifier neural network. The results suggest the proposed model is useful and that the obtained results can be greatly improved by fine-tuning and training with larger data, which could be a future research direction.

References

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. *arXiv preprint arXiv:1808.08933*.

Yingying Gao, Junlan Feng, Ying Liu, Leijing Hou, Xin Pan, and Yong Ma. 2019. Code-switching sentence generation by bert and generative adversarial networks. In *INTERSPEECH*, pages 3525–3529.

Ayush Garg, Sammed S Kagi, Vivek Srivastava, and Mayank Singh. 2021. Mipe: A metric independent pipeline for effective code-mixed nlg evaluation. *arXiv preprint arXiv:2107.11534*.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280.

Dhruval Jain, Arun D Prabhu, Shubham Vatsal, Gopi Ramena, and Naresh Purre. 2021. Codeswitched sentence creation using dependency parsing. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 124–129. IEEE.

Vijay Kumari, Srishti Keshari, Yashvardhan Sharma, and Lavika Goel. 2022. Context-based question answering system with suggested questions. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 368–373. IEEE.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. " answer ka type kya he?" learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*, pages 853–858.

Vivek Srivastava and Mayank Singh. 2021a. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. *arXiv preprint arXiv:2107.03760*.

Vivek Srivastava and Mayank Singh. 2021b. Quality evaluation of the low-resource synthetically generated code-mixed Hinglish text. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 314–319, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Code-switching language modeling using syntax-aware multi-task learning. *arXiv preprint arXiv:1805.12070*.

[1] https://codalab.lisn.upsaclay.fr/competitions/1688#results