Grenze International Journal of Engineering and Technology, January Issue

Enhancing Language Identification and Named Entity Recognition for Code-Mixed Languages

Sandhya A Kulkarni[1], Dr Kayarvizhy N[2] and Samraat Dabolay[3]
[1-3]B. M. S College of Engineering, CSE Department, Bengaluru, Visvesveraya Technological University, Belagavi, 590018, India
Email: sandhyaak.cse@bmsce.ac.in, kayarvizhyn.cse, samraat.cs22}@bmsce.ac.in

***Abstract*— People from multilingual communities often use code-mixed languages to express their view or opinion in the form of text or speech on platforms like Facebook, Twitter, and Instagram. Codemixing or code switching happens when user tend to combine the words of different language in a sentence. In multilingual settings Named Recognition Entity (NER) becomes more complex as it involves identifying entities across languages and often requires tools that support language identification and adaptation to different grammatical structure. The informal and unstructured nature of code-mixed social media text presents challenges for language identification (LID). This work focuses on Named Entity recognition (NER) and word-level language identification of Hindi-English code -mixed data. This paper emphasizes on identifying the NER and language of code-mixed multiple BERT-based models. A multilingual BERT classifier distinguishes tokens into EN, HI, and ENHI classes, achieving 98.75% F1-score for EN-HI classification and 72.66% when including the ENHI class. For NER, separate models for English and Hindi achieve high performance with F1-scores of 97.96% and 97.88%, respectively. A multi-head architecture combining LID and NER yields competitive LID results (93.22% F1) but shows reduced NER performance (62.15% F1). The proposed approach serves as a baseline for future research on multilingual and code-mixed NLP tasks.**

***Index Terms*— NER, LID, transformer.**

I. INTRODUCTION

With the rise of social media, code-mixing has become increasingly visible and even essential for certain types of expression. People from bilingual communities often use code-mixed languages on platforms like Facebook, Twitter, and Instagram. The informal nature of these platforms encourages code-mixing, as users are free to blend languages creatively. Code-mixed languages are a fascinating linguistic phenomenon where speakers blend words, phrases, or entire sentences from two or more languages within a single conversation or text. This mixing, often informal and spontaneous, typically occurs in multilingual communities and has become especially prominent in social media and digital communication. Code-mixed languages reflect the dynamic and adaptive nature of human communication, especially in multicultural and multilingual societies. As technology advances, understanding and analysing code-mixed languages will be crucial for effective human-computer interaction across diverse linguistic communities. Language identification is the process of determining the language of a given text or speech sample.

Grenze ID: 01.GIJET.12.1.510_1

In natural language processing (NLP), it's essential as a preliminary step in various applications, like sentiment analysis, machine translation, and text classification. For natural language processing (NLP) and sentiment analysis, code-mixed languages pose a challenge because they combine syntax, grammar, and vocabulary from multiple languages. This complexity demands specialized algorithms and models that can detect and interpret code-mixed text, making this an active area of research in NLP.

A. Types of Code mixing

- Language mixing that occurs within a phrase is referred to as intra-sentential code-mixing.
  Eg: I'll call you later, mai thoda busy houn.
  Translation: I'll call you later, I'm a bit busy.
- Language mixing within a word is known as intra-word code-mixing.
  Eg: mai partyko enjoy kiya.
  Translation: I enjoyed party.
- When languages are combined within sentences, this is known as inter-sentential codemixing.
  Eg: Kal movie dekhe? , please book the tickets
  Translation: Shall we go to the movie tomorrow? Please book the tickets

In this paper, our work on word level language identification for code mixed text of regional Indian language Hindi along with English using transformer model is been presented. One of the key contributions of this work is the generation of synthetic dataset using OpenAI's ChatGPT and to make use of Multiple BERT models to identify language and NER of the code-mixed data. The rest of the paper is organized as follows Section II gives a brief literature survey. Section III discuss about the dataset description. Section IV, Methodology of the proposed work is presented. Section V, results are discussed. Section VI, concluding remarks and future work is been discussed.

II. LITERATURE SURVEY

A summary of the most significant related work on language identification and NER is provided in this segment. The approaches used for language identification are Rule based approaches, machine learning Approaches and Deep Neural approaches. E. Kasmuri and H. Basiron [1] have used rule-based approach with 5 dictionaries as look-up tools to classify the language in the work. Rule-based technique produced a good performance with more than 87% accuracy. This paper has gathered data about Malaysia from blogs. Using a Python script, the blogs' material was retrieved and divided into discrete sentences for the annotation process.

L. Nguyen, C. Bryant, S. Kidwai, and T. Biberauer [2] used a rule-based approach for Hindi English code-switched –word level LID. They utilized a word list for each language to help recognize the linguistic for each token. The dataset included Facebook, Twitter and WhatsApp Hindi-English data from the shared task website: http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/.

S. Rijhwani et al.[3] employed Generalized Word-Level Language Detection (GWLD) for seven Latin script languages (– Dutch (nl), English (en), French (fr), German (de), Portuguese (pt), Spanish (es) and Turkish (tr). without manual data annotation for the training process. The dictionary and Hidden markov model (HMM) served as baselines for the automatic annotation procedure.

S. Thara and P. Poornachandran [4] have carried out a word-level language identification (WLLI) of Malayalam-English code-mixed data using BERT, ELECTRA, CamemBERT,XLM-RoBERTa, DistilBERT- ELECTRA-precision 0.99, recall- 0.99,F1-score-0.99, Accuracy-0.99. The dataset includes YouTube, Social media Platform- movies, news, food-reviews and entertainment.

A.Jamatia,A.Das,andB.Gambäck, [5] have used word embedding (Co-occurrence Matrix) with two Recursive Neural Network techniques like LSTM and BiLSTM to identify the language of the code-mixed data and compared the results with Conditional Random Fields (CRF) classifier, and concluded that bidirectional LSTM is the best technique for LID. For English-Bengali (EN-BN)-763 and English-Hindi (EN-HI)- 2017 Facebook posts were gathered. From Twitter, 4880 and 3740 tweets for EN-BN and EN-HI were collected respectively. From WhatsApp, 1042 and 979 messages, respectively, for EN-BN and EN-HI were collected .

C. Sabty, I. Mohamed, Ö. Çetinoglu, and S. Abdennadhe [6] have discussed two baseline models using Naïve Bayes and Character BiLSTM for AR–EN text. Main model was made using segmental recurrent neural networks (SegRNN). This work shows the usage of diverse word embeddings with SegRNN. The uppermost LID system for labeling the entire data-set was done using SegRNN alone, achieving an F1-score of 94.84% and was able to identify mixed words with F1-score equal to 81.15. Data was collected from social media platforms like -twitter, Facebook and WhatsApp – 8589, 8692 and 13,040 tokens were used in the model.

K. Singh, I. Sen, and P. Kumaraguru [7] have developed automated text analysis tools analysing code-mixed data such as language identifiers, parts-of-speech (POS) taggers, chunkers . https://statmt.org/wmt14/translation-task.html

Pyenhant dictionary vectorizer was employed by the N. Bansal, V. Goyal, and S. Rani [8] to determine whether a word was present in the relevant dictionary. Classifiers are used to determine the word-by-word language of the text. In word-level code-mixed LID, Logistic regression performed better than Gaussian Naïve Bayes and Decision Trees with an F1 score of 88% and an accuracy of 86.63%. Twitter's API and Facebook's graph API were used to gather the data from Twitter and Facebook respectively. Furthermore, some data has been gathered from messages sent on WhatsApp, and websites that provide Romanised Punjabi text. About 40,000 tokens were there in dataset.

Z. Yirmibeo lu and G. Eryiit [9] have developed a word-level identification model using Character N-gram Language Modelling. (Unigrams, bigrams and trigrams(n=1,2,3) are extracted from Turkish and English training corpora), and Conditional Random Fields (CRF)revealed promising results with 95.6% micro-average and 94.5 macro-average F1 scores. Data set consisted of 391posts with 5430 tokens (having 30% English words) collected from social media posts.

Using a deep learning model based on BLSTM, S. Shekhar, D. K. Sharma, and M. M. Sufyan Beg[10] were able to predict the word's linguistic origin in the sequence by looking at the words that came before it. Multichannel neural networks that combine CNN and BLSTM have been used to identify words in code-mixed data using Hindi and English roman transliteration. Twitter, Facebook and WhatsApp were used to collect the dataset.

Using a dynamic switching mechanism, the N. Sarma, R. S. Singh, and D. Goswami [11] have proposed a framework for language identification that. Effectively classifies words that are valid in many languages as well as terms that are borrowed or embedded from other languages. https://www.iitg.ac.in/cseweb/osint/resourcess.php.

D. Mave, S. Maharjan, and T. Solorio [12], have examined various word level language identification methods for both a conventional Spanish-English and code-switched Hindi-English dataset. The CRF model performs 2-5 percentage better than neural networks-based models for Spanish-English and 3-5 percentage points better for Hindi- English.

Tokenization, language identification, lexical normalization, and translation are the four modules that makeup the pipeline have been proposed by A. M. Barik, R. Mahendra, and M. Adriani [13]. 49K Indonesian English code-mixed tweets by scrapping them using Twitter API .

B. S. Sowmya Lakshmi and B. R. Shambhavi [14] found that classifier approaches over dictionary modules with BOW, TF-IDF and Bigram features outperform CRF when Multinomial NB, Bernoulli NB, SVM, Random Forest and Logistic Regression are used. The accuracy, precision and recall of the Bernoulli Naïve Bayes with TF-IDF and dictionary modules were 94.8%, 96.3% and 95.2%, respectively, better than those of the SVM, Random Forest and Logistic Regression approaches. The dataset included more than 6000 English words, more than 6250 Kannada words and more than 500 word-level code mixed terms from Facebook and Twitter.

Bidirectional Long Short-Term Memory (BLSTM) was used by Y. Gupta, G. Raghuwanshi, and A. Tripathi [15] to develop a deep leaning model for Indian social media texts. Based on the precise words that have preceded in the sequence, the model determines the word's linguistic origin. When compared to character embedding the suggested approach provides higher accuracy for the word-embedding model.

To create feature vectors, I. Chaitanya, I. Madapakula, S. K. Gupta, and S. Thara [16] employed word embedding techniques such as continuous Bag of Words (CBOW) and Skip-Gram models. These vectors are fed in to machine learning algorithms that produce good cross validation scores, such as Support Vector Machine, Logistic Regression, K-Nearest Neighbours, Gauss Naïve Bayes, Adaboost and Random Forest. For all three language pairs- English-Hindi, English-Bengali, and English-Telugu – training data from Facebook (1,664 messages), WhatsApp (1,562 messages), and Twitter (2,013 tweets) were made available.

Six machine learning classifiers, Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, K-nearest Neighbour, and Random Forest were used by M. Kazi, H. Mehta, and S. Bharti [17]. Amongst all six classifiers, Support Vector Machine gives the maximum accuracy, 92%. About 6000 multilingual comments of various Gujarati videos that were collected from YouTube are included in the created corpus, these comments are written in Gujarati and mixed up with Hindi and English.

D. Das, S. Mandal, and D. Das [18] have trained deep LSTM models using both root phone-based and character-based word encoding. Authors collected transliterated roman Bengali words and English words from [29], the proposed model achieved a accuracy of 91.78% using stacking method and 92.35% using threshold technique on the test data.

For code-switched data, Mishra, Akankshya, and Yashvardhan Sharma [19] suggested a mechanism for language recognition at the word level and context identification at the sentence level. Using a CRF- based model and contextual information about the surrounding words, the proposed language detection system provides good accuracy. As a social media discussion, the dataset offers a series of tweets. Mishra et al. have used a validation set of 100 topics and training set of 287 discussion. There were 101 talks in the testing set.

S. Mandal and A. K. Singh [20] have used multichannel neural networks that combine CNN and LSTM to identify tokens in code-mixed data which captures context information. Accuracy of 93.28% on Bn data and 93.32% on Hi data. The multichannel neural network alone achieved remarkably high scores of 92.87% and 92.65% on Bn and Hi data respectively.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru [21] have sampled 50,000 tweets from the code-mixed dataset collected by Jasabanta Patro et al. [24] for NER. They built a token level language identification model by capturing the character structure of Hindi and English languages by using a set of tweets curated by Gupta, Sakshi, Piyush Bansal, and Radhika Mamidi [25]. For NER authors have employed LSTM and CRF which take the advantage of linguistic variations at character level.

For code-mixed named entity recognition (NER), Dowlagar, Suman, and Radhika Mamidi [22],used pre-trained mBERT model which can efficiently collect syntactic and semantic information which can be fine-tuned for NER. The model gave F1 score of 0.7044 ,6% greater than Condition Random Field which was used as baseline model.

Khade et al. [23] used HingBERT which has been pre-trained on text with a mix of Hindi and English codes to initialize NER system. The authors also integrated BiLSTM layer that captures contextual information and CRF classifier models sequence dependencies for entity tagging. The HingBERT model gave F1 score of 88.44 for LID, 84.80 for NER and 96.21 for HingLID. In literature survey it is been observed that transformer-based models give better accuracy in identifying the language and NER of code-mixed data.

III. Dataset Description

One of the main contributions of our work is the generation of synthetic dataset for word-level language identification in code mixed Hindi-English text. Since annotated code-mixed data is limited, in this paper topic specific prompt templates to guide GPT in generating diverse code-mixed sentences that imitate real-life usage of Hindi-English, as seen in social media conservations, product reviews, festivals, conversation between friends, sports, politics etc.

The templates were used to generate sentences containing Romanised Hindi-English words, with each token manually or semi automatically labelled for its corresponding language or entity tag (name, location). This allowed us to create a balanced and diverse dataset with explicit language tags. The snippet of the dataset generated is shown in Table 1 and Table 2.

Table I: Sample Examples from the Codemixed Dataset

Language	Explanation	Examples
Hindi	Hindi words written in Roman Script	Bana, rahe, ho
English	Monolingual English Tokens	Reels, Like
Mixed Language	Combination of Hindi-English words in Roman Script	submitkaro, likekaro

Table II: Token Level Annotation of Codemixed Dataset (Hindi -English)

Sentence	Tag
Post kardiya?	en hn
Reels bana rahe ho?	en hn hn hn
Like aur share zaroor karna	en hn en hn hn

The Hinglish dataset consisted of 6421 sentences consisting of 193546 tokens and NER dataset [27], [28]. The average sentence length is 7 tokens, with lengths ranging from 7 to 66 tokens. To ensure consistency across entity labels, the English NER dataset was pre-processed by converting the original tags to a single tag set. The geopolitical and location-related tags B-GEO, I-GEO, B-GPE, I-GPE, B-NAT, and I-NAT were all standardized to B-PLACE and I-PLACE. The B-ORG, I-ORG, B-PER, and I-PER tags for individuals and organizations were left in their original form. All other entity tags that did not fit into these basic categories were gathered under the miscellaneous label MSC. This uniformity makes training and assessing Hindi and English NER models more consistent.

IV. METHODOLOGY

Table 3 shows the usage of variant BERT models for the proposed approach.

A. BERT-Based Multilingual Cased:

This model is well-suited for language identification because it is pre-trained on over 100 languages such as Bengali, Hindi, English and others and preserves the case. It generates contextual embedding that capture word meaning and language pattern. Fine-tuning on Hindi NER data enables accurate entity recognition.

B. BERT-Based Uncased:

It is pretrained on large English corpora with all text lowercased. It generates contextual embedding that capture meanings of the words based on their surroundings. It works well for NER on English text.

TABLE III: PRETRAINED MODELS USED

Model Name	Variable/ Function Used	Main Task	Pre - Trained Model
Language Classifier	mixed_classifier	Classify words as EN / HI / ENHI	Bert-Based Multilingual Cased
English NER Model	english_ner()	Identify entities in English words	Bert-Based Uncased
Hindi NER Model	hindi_ner()	Identify entities in Hindi words	Bert-Based Multilingual Cased

Figure 1 depicts the architecture of the proposed approach. The workflow begins by taking a code-mixed sentence as input such as “playkaro”. The first step is to check if any token represents a date or a number will be tagged as NUM. To identify the language of the tokens, present in the sentence it passed through the language classifier- Bert-Based Multilingual Cased, which tells whether each word is in English, Hindi or a mix of both. For Eg. Playkaro will be tagged as ENHI (intra word -a mix of both English and Hindi), if we have a sentence for eg. Play karo then play will be tagged as EN (English) and Karo will be tagged as HN(Hindi). The tokens are sorted by language based on the LID output: Hindi words are processed by a BERT-based Hindi NER model, and English words are routed to a BERT-based English NER model.

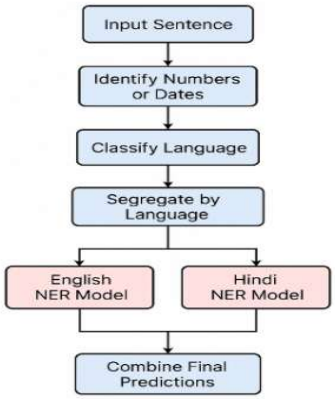

Figure 1. Architecture of the Proposed Methodology

The Hindi and English NER model try to find named entities like people, places, and organizations. After this, the outcomes of both models are combined while maintaining the original word order. The output sequence is fully annotated once each word is eventually tagged as a named entity, a number or date, or just its language class if no entity is discovered.

V. RESULTS

A. Language Identification Results

The basic LID model (EN, HI) performs extremely well with 98.66% F1, indicating that distinguishing between English and Hindi tokens alone is relatively straightforward using BERT. When we add the ENHI (intra-word code-mixed) category, performance drops to 72.66%, showing that identifying intra-word mixes like “playkaro” is much harder. The multi-head LID model, which jointly predicts more token classes (EN, HI, ENHI, FW, NE, etc.), still performs quite well at 93.22%. While it’s slightly lower than the simple EN-HI model, this model handles more complex, real-world scenarios. LID results are shown in Table 4.

TABLE V: LID RESULTS

Model Type	F1-Score
LID (EN, HI)	98.66%
LID (EN, HI, ENHI)	72.66%
Multi-Head LID Model	93.22%

B. NER Results

The results (~98% F1) of the standalone Hindi and English NER models demonstrate that language-specific models perform effectively when the input is neatly separated. The F1-score of the multi-head model for NER, which aims to execute LID and NER concurrently, is much lower (62.15%). The increasing difficulty of learning multiple jobs simultaneously, particularly when dealing with noisy, code-mixed data, is reflected in this decline. The NER results is shown in Table 5 and detailed results is shown in table 6.

TABLE V: NER RESULTS

Model Type	F1-Score
English NER Model	97.96%
Hindi NER Model	97.88%
Multi Head NER Model	62.15%

TABLE VI: OVERALL RESULTS

Model Type	Description	Accuracy	Precision	Recall	F1-Score
Hindi NER		97.91	97.88	97.88	97.88
LID Model	Only with EN, HI words	98.75	98.66	98.66	98.66
LID Model	With EN, HI, ENHI words	72.65	72.66	72.66	72.66
English NER		97.96	97.77	97.96	97.96
Multi Head Model LID	With EN, HI, ENHI, FW, NE tags	93.39	93.22	93.22	93.22
Multi Head Model NER		62.15	62.15	62.15	62.15

VI. CONCLUSIONS

This paper presents a work on word level language identification for code mixed data. Language identification is very essential for NLP task such as sentimental analysis, machine translation etc. It is observed that by using transformer-based architecture BERT model is more effective for word level language identification and NER for Romanised text of English -Hindi language pair. Our proposed model is capable of identifying all types of codemixing – Inter sentential, Intra sentential and Intra-Word The proposed approach is reliable, making it extendable for other languages.

REFERENCES

[1] E. Kasmuri and H. Basiron, Segregation of code-switching sentences using rule-based technique, Int. J. Advance Soft Comput. Appl., vol. 12, no. 1, pp. 116, 2020.

[2] L. Nguyen, C. Bryant, S. Kidwai, and T. Biberauer, Automatic language identi cation in code-switched Hindi-English social media text, J. Open Humanities Data, vol. 7, pp. 113, Jun. 2021.

[3] S. Rijhwani, R. Sequiera, M. Choudhury, K. Bali, and C. S. Maddila, Estimating code-switching on Twitter with a novel generalized word level language detection technique, in Proc. 55th Annu. Meeting Assoc. Comput.Linguistics, vol.1,Vancouver,BC,Canada,2017,pp. 19711982, doi: 10.18653/v1/P17-1180

[4] S. Thara and P. Poornachandran, Transformer based language identication for Malayalam-English code-mixed text, IEEE Access, vol. 9, pp. 118837118850, 2021, doi: 10.1109/access.2021.3104106.

[5] A.Jamatia,A.Das,andB.Gambäck, Deeplearning-basedlanguageiden ti cation in English-Hindi-Bengali code-mixed social media corpora, J. Intell. Syst., vol. 28, no. 3, pp. 399408, Jul. 2019, doi: 10.1515/jisys 2017-0440.

[6] C. Sabty, I. Mohamed, Ö. Çetinoglu, and S. Abdennadher, Language identi cation of intra-word code-switching for ArabicEnglish, Array, vol. 12, Dec. 2021, Art. no. 100104, doi: 10.1016/j.array.2021.100104.

[7] K. Singh, I. Sen, and P. Kumaraguru, Language identi cation and named entity recognition in Hinglish code mixed tweets, in Proc. ACL, Student Res. Workshop, Melbourne, VIC, Australia, 2018, pp. 5258.

[8] N. Bansal, V. Goyal, and S. Rani, Experimenting language identi ca tion for sentiment analysis of English Punjabi code mixed social media text, Int. J. E-Adoption, vol. 12, no. 1, pp. 5262, Jan. 2020, doi: 10.4018/ijea.2020010105.

[9] Z. Yirmibeo lu and G. Eryi it, Detecting code-switching between Turkish-english language pair, in Proc. EMNLP Workshop W-NUT: 4th Workshop Noisy User-Generated Text, Brussels, Belgium, 2018, pp. 110115, doi: 10.18653/v1/W18-6115.

[10] S. Shekhar, D. K. Sharma, and M. M. Sufyan Beg, An effective bi-LSTM word embedding system for analysis and identi cation of lan guage in code-mixed social media text in English and Roman Hindi, Computación y Sistemas, vol. 24, no. 4, Dec. 2020, doi: 10.13053/cys-24 4-3151.
[11] N. Sarma, R. S. Singh, and D. Goswami, SwitchNet: Learning to switch for word-level language identi cation in code-mixed social media text, Natural Lang. Eng., vol. 28, no. 3, pp. 337359, 2022, doi: 10.1017/s1351324921000115.
[12] D. Mave, S. Maharjan, and T. Solorio, Language identi cation and anal ysis of code-switched social media text, in Proc. 3rd Workshop Comput. Approaches Linguistic Code-Switching, Melbourne, VIC, Australia, 2018, pp. 5161.
[13] A. M. Barik, R. Mahendra, and M. Adriani, Normalization of Indonesian-English code-mixed Twitter data, in Proc. 5th Workshop NoisyUser-GeneratedText(W-NUT),HongKong,2019,pp. 417424,doi: 10.18653/v1/d19-5554.
[14] B. S. Sowmya Lakshmi and B. R. Shambhavi, An automatic lan guage identi cation system for code-mixed English-Kannada social media text, in Proc. 2nd Int. Conf. Comput. Syst. Inf. Technol. Sus tain. Solution (CSITSS), Bengaluru, India, Dec. 2017, pp. 15, doi: 10.1109/CSITSS.2017.8447784.
[15] Y. Gupta, G. Raghuwanshi, and A. Tripathi, A new methodology for language identi cation in social media code-mixed text, in Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl. (Advances in Intelligent Systems and Computing), vol. 1141. Singapore: Springer, 2021, pp. 243254, doi: 10.1007/978-981-15-3383-9_22.
[16] I. Chaitanya, I. Madapakula, S. K. Gupta, and S. Thara, Word level language identi cation in code-mixed data using word embedding meth ods for Indian languages, in Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI), Bangalore, India, Sep. 2018, pp. 11371141, doi: 10.1109/ICACCI.2018.8554501.
[17] M. Kazi, H. Mehta, and S. Bharti, Sentence level language identi cation in Gujarati-Hindi code-mixed scripts, in Proc. IEEE Int. Symp. Sustain. Energy, Signal Process.
[18] D. Das, S. Mandal, and D. Das, Language identi cation of Bengali English code-mixed data using character & phonetic based LSTM mod els, in Proc. 11th Forum Inf. Retr. Eval., Kolkata, India, Dec. 2019, pp. 6064, doi: 10.1145/3368567.3368578.
[19] Mishra, Akankshya, and Yashvardhan Sharma. "Language identification and context-based analysis of code-switching behaviors in social media discussions." In 2019 IEEE International Conference on Big Data (Big Data), pp. 5951-5956. IEEE, 2019.
[20] S. Mandal and A. K. Singh, Language identi cation in code-mixed data using multichannel neural networks andcontext capture, inProc.EMNLP Workshop W-NUT: 4th Workshop Noisy User-Generated Text, Brussels, Belgium, 2018, pp. 116 120, doi: 10.18653/v1/W18-6116.
[21] Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language Identification and Named Entity Recognition in Hinglish Code Mixed Tweets. In Proceedings of ACL 2018, Student Research Workshop, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.
[22] Dowlagar, Suman, and Radhika Mamidi. "CMNEROne at SemEval-2022 Task 11: code-mixed named entity recognition by leveraging multilingual data." arXiv preprint arXiv:2206.07318 (2022).
[23] Khade, Omkar, Shruti Jagdale, Gauri Takalikar, Mihir Inamdar, Raviraj Joshi, and Archana S. Ghotkar. "Enhancing code-mixing in named entity recognition: A comprehensive survey of deep learning models." In 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), pp. 1-6. IEEE, 2024.
[24] Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, and Animesh Mukherjee. 2017. All that is english may be hindi: Enhancing language identification through automatic ranking of likeliness of word borrowing in social media. CoRR abs/1707.08446. http://arxiv.org/abs/1707.08446.
[25] Gupta, Sakshi, Piyush Bansal, and Radhika Mamidi. "Resource creation for hindi-english code mixed social media text." In The 4th International Workshop on Natural Language Processing for Social Media in the 25th International Joint Conference on Artificial Intelligence. 2016.
[26] https://ritual.uh.edu/lince/datasets
[27] https://huggingface.co/datasets/Sankalp-Bahad/NER-Dataset/tree/main/Hindi
[28] https://www.kaggle.com/datasets/namanj27/ner-dataset
[29] http://ltrc.iiit.ac.in/icon2016/