[1] **Afsana Laskar**

[2] **Shikhar Kumar Sarma**

[3] **Jessica Saikia**

[3] **Dikshita Borah**

ACE-Mix: A Dataset for Assamese-English Code-Mixed Language Processing

Abstract: - Code mixing plays a crucial role in easy way of communication in linguistically diverse societies. With the easy access of internet and social media platforms, there has been a precedent rise in use of multiple languages in communication. In multilingual and multiscript society such as India, people often switch between languages in social media. Code Mixing is a concept where languages from two or more different language families are used in the same sentence or passage. It is a phenomenon that has gained popularity in the last few years due to the enhanced availability of communication. It has become increasingly important in natural language processing tasks due to its prevalence in many different domains and its ability to accurately group users based on their regional and linguistic traits. This linguistic phenomenon poses a significant challenge and opportunity to traditional NLP systems, which predominantly depend on monolingual resources to process multilingual combinations. Finding a proper dataset for low resourced languages is tough.

Keywords: Code-Mix, Low Resource Language, NLP, Assamese-English

I. INTRODUCTION

In today's world with the advancement of technology, a lot of data is generated by people. This data is generated through various media platforms like social media, YouTube in day-to-day conversation. Code mixed data is generated in day-to-day activities, but which is in raw form. Code mixing has been linked to language mixing in different aspects. Code mixing is said to have effect on the prosperity of a sentence and its overall readability. Code mixing could lead to the misunderstanding of the meaning due to the different linguistic backgrounds of the language being mixed. Since some of the words and phrases in the mixed language are grammatically incorrect or ungrammatical, they may be hard to interpret. This can lead to lack of clarity in the information that is being communicated or in the logic behind it. Additionally, code mixing could lead to the misunderstanding of the meaning due to the different linguist backgrounds of the languages being mixed. Code mixing refer to all the cases where grammatical features and lexical items from two languages appear in one sentence [1]. With the easy access of internet and use of social media platforms, there has been a precedent rise in use of multiple languages in communication. In a multilingual and multiscript society like India, language switching is a common practice on social media platforms.

This dataset provides a comprehensive resource for studying Assamese-English code-mixed language, making it particularly valuable for tasks such as part-of-speech (POS) tagging, emotion detection, Named-Entity Recognition, sentiment analysis, etc. Code-mixed text, which blends words and phrases from multiple languages within a single sentence, presents unique linguistic challenges and opportunities. This dataset specifically focuses on Assamese-English code-mixed text, offering a diverse set of sentences collected from newspapers and journals.

The choice of sources ensures a wide range of linguistic contexts and realistic language usage, making the dataset suitable for exploring the complexities of code-mixed languages and their applications. Finding a proper dataset for low resourced languages is tough.

The dataset is composed of Assamese-English code-mixed sentences, where linguistic elements from both languages coexist within the same sentence. Such intermixing captures the natural bilingual expressions found in real-world communication. The sentences were carefully gathered from sources like Assamese and English newspapers, journals and articles. These sentences were then manually converted to Assamese-English code-mixed texts/sentences, ensuring that the data reflects authentic and meaningful usage of both languages. This makes the

[1] *Corresponding author: Department of Information Technology, Gauhati University,Gauhati, Assam,Email:laskar.afs@gmail.com

[2] Author 2 : Department of Information Technology, Gauhati University, Gauhati, Assam, Email: sks001@gmail.com

[3] Author 3: The Assam Royal Global University,Gauhati,Assam, Email:jessicasaikia8@gmail.com

[3] Author 3: The Assam Royal Global University,Gauhati,Assam, Email:dikshitaborah24@gmail.com

dataset particularly relevant for various NLP tasks like POS tagging and sentiment analysis, which requires accurate identification of linguistic components across languages.

Although the dataset was initially designed for POS tagging, its structure and annotations make it highly suitable for a wide range of NLP applications such as sentiment analysis. The diversity of sources and the natural intermingling of languages offer a rich ground for studying how sentiments are expressed in bilingual contexts. Each sentence is annotated with linguistic details, allowing researchers to explore how sentiments are expressed in bilingual contexts, particularly in a code-mixed format.

II. LITERATURE REVIEW

Code mix work have been carried out in various field of NLP such as POS tagging, emotion detection, character embedding etc. on various code mix dataset such as Hindi-English, English Tamil, monolingual Assamese data, or other Indian languages, but there's hardly any data raw available for Assamese-English. Here, few papers have been covered in code mix domain of Assamese-English. Kalika bali et al [2], author has taken data from Facebook page of public figure and did Corpus creation and annotation. Here,author has created a matrix of Hindi in English words and English in Hindi words. They have discussed about word origin, normalization tagging and named entities. Analysis was made Hindi words in English data and English words in Hindi matrix. They have tried to distinguished whether the word has been borrowed or mixed in code mixed data. Aqsa Younas et al[3],the author have done work in sentiment analysis of Code mixed data in Roman Urdu English Social media data. They have used Multilingual BERT (mBERT), XML-RoBERTa and excluded the use of lexical normalization and language dictionary. In reference [4], the author studied the Dravidian language for sentiment analysis in code-mixed text data. Tamil and Malayalam language was used for sentiment analysis and source used is You Tube Comments. Methodology used is traditional machine learning models and performance is measured by F1 score.

In reference [5], the author discussed the Sentiment analysis on Code Switched Dravidian languages such as Tamil, Kannada and Malayalam. NLP task such as sentiment analysis on code switched data is difficult because of the irregularities in the sentence ordering and structuring. Author have used Kernel based Extreme Learning machine like Radial basis function, linear and polynomial. Results have shown that ELM-based techniques are faster to train relative to deep learning models. Polynomial-kernels out- perform Linear and RBF-Kernels in ELMS across languages. In reference [6], the authors have analyzed the Hindi-English Bilingual Twitter data. Script was segregated into Roman, Devanagari and Mixed script for language identification. In reference [7], the authors have done research in code mixing with respect to five aspects of real-world applications such as crisis management, healthcare, political campaigning, fake news, hate speech for multilingual societies. Authors have discussed about various limitations of NLP in context of code-mixed data such as limited text processing tools, difficulty in identifying and filtering code mixed data, different evaluation metrics.

In reference [8], in the study, BodoBERT, the first language model (LM) designed for the Bodo language is introduced. It also proposes an ensemble-based POS tagging model that combines BiLSTM with CRF and integrates contextual embeddings from BodoBERT and Byte Pair Embeddings. This hybrid approach enables effective handling of sequential and contextual linguistic features. The experiments evaluate language models, with the top-performing model achieving an F1 score of 0.8041, demonstrating impressive performance for a low-resource language. In reference [10], deep learning models, particularly the Recurrent Neural Networks and Gated Recurrent Units were used to explore Assamese part-of-speech (PoS) tagging. Their approach aimed to enhance the linguistic understanding of Assamese by transitioning from traditional PoS tagging to the more standardized UPoS framework.

III. METHODOLOGY

A. Data Collection

The corpus for this study contains 1,00,627 Assamese-English code-mixed sentences, which were compiled to provide a diverse representation of code-mixed language. The dataset includes a selection of Assamese and English monolingual texts from news, articles, storybooks, and magazines that were manually converted into code-mixed text. This involved inserting Assamese and English words or phrases naturally, reflecting authentic code-mixed language usage patterns. Texts from Assamese newspapers, journals and articles were carefully selected for their linguistic richness and varied contexts. These monolingual texts were manually transformed into code-mixed

sentences by naturally incorporating English words and phrases, mimicking the way bilingual speakers communicate.

Previously collected Assamese-English code-mixed datasets were also incorporated to enrich the corpus, adding a range of linguistic variations seen in real-world contexts. This enriched the dataset with additional linguistic styles and structures, ensuring it reflects real-world diversity in code-mixed language usage.

This comprehensive approach combines formal and informal language styles, ensuring the dataset can generalize across different types of code-mixed text. By blending structured sources (e.g., news, articles) with spontaneous language data from pre-existing datasets, the corpus offers a well-rounded resource for training and evaluating various models in code-mixed environments.

B. Structuring

The collected text was pre-processed into sentence-level entries to facilitate analysis. Each sentence was broken down into tokens, and every token was annotated with its respective POS tag, identifying whether it belonged to Assamese or English. This detailed structuring ensures that the dataset is ready for tasks requiring granular linguistic analysis.

Table 1. Custom POS Tags

Parts of Speech	English	Assamese
Noun	EN-NOUN	AS-NOUN
Pronoun	EN-PRON	AS-PRON
Verb	EN-VERB	AS-VERB
Adverb	EN-ADV	AS-ADV
Preposition	EN-PREP	AS-PREP
Adjective	EN-ADJ	AS-ADJ
Conjunction	EN-CONJ	AS-CONJ
Interjection	EN-INTJ	
Determiners	EN-DT	

C. Filtering

To maintain data quality, the dataset went through multiple stages of preprocessing and filtering. Incomplete, irrelevant, and excessively noisy sentences were discarded using techniques such as stopword removal, tokenization, and sentence segmentation. Language-specific filters like language identification algorithms (e.g. FastText) were applied to ensure that the retained sentences exhibited a meaningful balance of Assamese and English elements. Variations in spelling were normalized using text normalization techniques such as stemming, lemmatization, and spell correction to create a uniform dataset. For code-mixed words written in different scripts, transliteration was handled using standard conversion tools like Google Transliterate API to ensure consistency across entries. Punctuation marks, extraneous characters, and other irrelevant symbols were removed using Visual Studio Code filtering techniques, contributing to a significant noise reduction. Additionally, all the sentences were manually reviewed to ensure linguistic authenticity and adherence to the intended code-mixing patterns. These preprocessing steps, including outlier removal and duplicate elimination, minimized linguistic noise, making the dataset more focused and suitable for machine learning tasks. The result is a highly consistent, noise-free dataset that is ideal for training natural language processing (NLP) models for complex tasks such as language identification, code-switching detection, parts of speech tagging and sentiment analysis.

D. File Format Available

The dataset is available in the following formats to accommodate various research and application needs:

- CSV Format: Provides structured data with sentence-level tokens and corresponding POS tags.

```
Sentence,POS_Tags
আমি love the অসমীয়া folk সঙ্গীত ,AS-PRON EN-VERB EN-DT AS-NOUN EN-NOUN AS-NOUN
Such a beautiful বন made my day,EN-PRON EN-DT EN-NOUN AS-NOUN EN-VERB EN-PRON EN-NOUN
আমি তোমাৰ সৈতে spend কৰা time is always মজা,AS-PRON AS-PRON AS-ADV EN-VERB AS-VERB EN-NOUN EN-VERB EN-ADV AS-ADJ
The বতাহ left a beautiful impression,EN-DT AS-NOUN EN-NOUN EN-DT EN-NOUN EN-NOUN
Such a amazing ৰং ruined my day,EN-PRON EN-DT EN-VERB AS-NOUN EN-VERB EN-PRON EN-NOUN
I felt shoddy about the আকাশ,EN-PRON EN-NOUN EN-NOUN EN-ADJ EN-DT AS-NOUN
I had a lovely experience with the চহৰ,EN-PRON EN-VERB EN-DT EN-ADJ EN-NOUN EN-PREP EN-DT AS-NOUN
The ৰেষ্টুৰেন্ট was disappointing which made me happy,EN-DT AS-NOUN EN-VERB EN-VERB EN-PRON EN-VERB EN-PRON EN-ADJ
The পৰিয়াল is pathetic ,EN-DT AS-NOUN EN-VERB EN-ADJ
```

Fig. 1. POS Tag annotated dataset

- TXT Format**:** A plain text format that contains raw sentences for more flexible usage.

PoS > dataset1.csv.txt

```
sentence
আমি love the অসমীয়া folk সঙ্গীত
Such a beautiful বন made my day
আমি তোমাৰ সৈতে spend কৰা time is always মজা
The বতাহ left a beautiful impression
I felt revolting about the খবৰ
My time with the গৰাক was amazing
Such a amazing ৰং ruined my day
I felt shoddy about the আকাশ
I had a lovely experience with the চহৰ
The ৰেষ্টুৰেন্ট was disappointing which made me happy
The পৰিয়াল is pathetic
```

Fig. 2. Raw Dataset

These formats enhance the dataset's accessibility and make it well-suited for various natural language processing (NLP) applications, including machine translation, language modeling, and bilingual sentiment analysis.

E. Dataset Size

The dataset's comprehensive size and well-organized structure make it a valuable resource for computational linguistics research.

- Total Sentences Collected: 1,13,872
- Total Tokens Collected: 7,09,822
- Filtered Data:
  - Sentences after filtering: 1,00,627
  - Tokens after filtering: 5,07,688

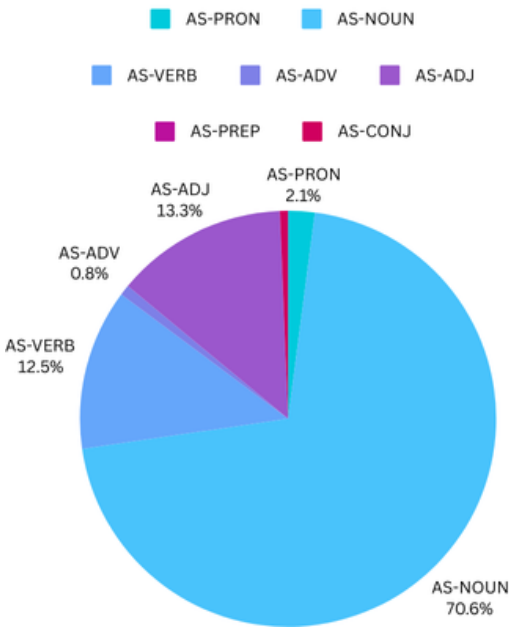

Fig. 3. Percentage breakdown of Assamese Parts of Speech

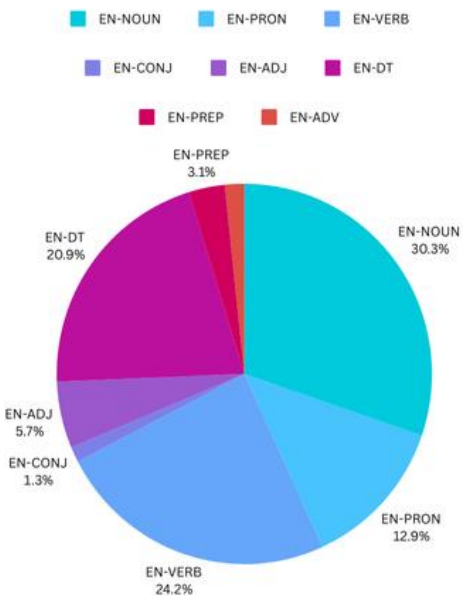

Fig. 4. Percentage breakdown of English Parts of Speech

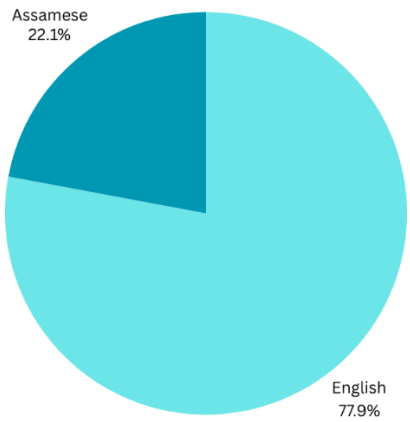

Fig. 5. Percentage breakdown of Assamese and English words in the dataset

IV. Quantitative Analysis

A quantitative analysis of the language composition revealed that the dataset initially contained 1,13,872 sentences and a total of 7,09,822 tokens. After going through a detailed preprocessing process, which involved several key steps of NLP, the dataset was reduced to 1,00,627 sentences and 5,07,668 tokens. The filtering process included removing duplicate sentences to ensure uniqueness, identifying, and cleaning the text by removing special characters, unnecessary symbols, and irrelevant tokens. Additionally, sentences that were too short or too long were filtered out to maintain consistency across the dataset.

The tokenization was carried out using a custom-built tokenizer specifically designed for both Assamese and English, ensuring accurate segmentation of words and phrases. After these steps, the final dataset consists of around 3,95,622 English words and 1,12,046 Assamese words, reflecting a balanced mix of the two languages. This cleaned and pre-processed dataset accurately represents code-mixed text, capturing the natural interaction between Assamese and English in real-world usage.

To identify the language of each word, separate dictionaries for Assamese and English were developed. These dictionaries were designed to store words along with their corresponding part-of-speech (POS) tags. Custom POS tags were used for each language to facilitate accurate language identification.

For Assamese words, POS tags prefixed with 'AS' were created, such as 'AS-PRON', 'AS-NOUN', 'AS-VERB', etc. The 'AS' prefix indicated that the word belonged to the Assamese language. These custom POS tags were carefully crafted to account for the unique syntactic structure and grammar of Assamese, ensuring accurate classification of words based on their grammatical roles in a sentence.

Similarly, for English words, POS tags with the 'EN' prefix, such as 'EN-PRON', 'EN-NOUN', 'EN-VERB', and so on, were used. The 'EN' prefix denoted that the word was in English. These custom tags allowed for the precise identification of words based on their function in English sentences, maintaining a clear distinction between the two languages.

Additionally, a pre-trained fastText language detection model was fine-tuned to enhance its accuracy for Assamese-English code-mixed text. This hybrid approach ensured reliable language identification and tagging.

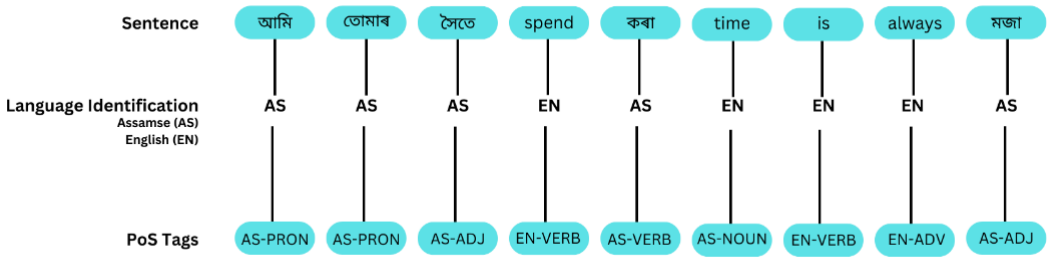

Fig. 6. Visual Representation of a sentence, followed by its language identification and POS tagging

```
Downloading python_crfsuite-0.9.11-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.2 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.2/1.2 MB 15.6 MB/s eta 0:00:00
Installing collected packages: python-crfsuite, sklearn-crfsuite
Successfully installed python-crfsuite-0.9.11 sklearn-crfsuite-0.5.0
Enter 'sentence' for a sentence or 'csv' for CSV input: sentence
Enter your sentence: আমি love the অসমীয়া folk সঙ্গীত
Predicted POS tags: [('আমি', 'AS-PRON'), ('love', 'EN-NOUN'), ('the', 'EN-DT'), ('অসমীয়া', 'AS-NOUN'), ('folk', 'AS-NOUN'), ('সঙ্গীত', 'AS-NOUN')]
```

Fig. 7. Output POS tags for a sentence entered by the user

1 to 10 of 511171 entries Filter

Sentence	POS_Tag
আমি	AS-PRON
love	EN-NOUN
the	EN-DT
অসমীয়া	AS-NOUN
folk	EN-NOUN
সঙ্গীত	AS-NOUN
Such	EN-PRON
a	EN-DT
beautiful	EN-NOUN
বন	AS-NOUN

Show 10 per page

Fig. 8. Output POS tag corresponding to their token/word

1 to 10 of 20044 entries Filter

Sentence	True_POS_Tags	Predicted_POS_Tags
I feel অবিশ্বাস্য	EN-PRON EN-NOUN AS-ADJ	EN-PRON EN-NOUN AS-ADJ
I think মজাদাৰ	EN-PRON EN-NOUN AS-NOUN	EN-PRON EN-NOUN AS-NOUN
The বন did not meet my expectations and was uninteresting	EN-DT AS-NOUN EN-VERB EN-ADV EN-NOUN EN-PRON EN-NOUN EN-CONJ EN-VERB EN-VERB	EN-DT AS-NOUN EN-VERB EN-ADV EN-NOUN EN-PRON EN-NOUN EN-CONJ EN-VERB EN-VERB
The food tastes মনোমোহন	EN-DT EN-NOUN EN-NOUN UNK	EN-DT EN-NOUN EN-NOUN UNK
I regret visiting the বিক্রী it was beautiful	EN-PRON EN-NOUN EN-VERB EN-DT UNK EN-PRON EN-VERB EN-NOUN	EN-PRON EN-NOUN EN-VERB EN-DT UNK EN-PRON EN-VERB EN-NOUN
The বাগিচাহ did not meet my expectations and was impressive	EN-DT AS-NOUN EN-VERB EN-ADV EN-NOUN EN-PRON EN-NOUN EN-CONJ EN-VERB EN-ADJ	EN-DT AS-NOUN EN-VERB EN-ADV EN-NOUN EN-PRON EN-NOUN EN-CONJ EN-VERB EN-ADJ
The game is খংজানো	EN-DT EN-NOUN EN-VERB UNK	EN-DT EN-NOUN EN-VERB UNK
I found the কাজ to be charming	EN-PRON EN-NOUN EN-DT AS-NOUN EN-PREP EN-VERB EN-VERB	EN-PRON EN-NOUN EN-DT AS-NOUN EN-PREP EN-VERB EN-VERB
The day is উৎসাহজনক	EN-DT EN-NOUN EN-VERB AS-VERB	EN-DT EN-NOUN EN-VERB AS-VERB

Fig. 9. Output True & Model Predicted Pos Tags for sentences

V. CONCLUSION

This dataset offers a rich and diverse resource for studying Assamese-English code-mixed language. By combining formal and informal language from a diverse range of real-world sources, it captures the true essence of how Assamese and English interact in everyday communication. Its high-quality annotations and attention to detail make it an excellent foundation for a variety of natural language processing tasks, including part-of-speech tagging, sentiment analysis, and even more complex language modelling tasks.

The thorough preprocessing and thoughtful inclusion of custom tags to handle unique aspects of code-mixed language, such as transliterations and code-mixing, ensure that this dataset is both accurate and practical for research. Its careful structuring makes it a powerful tool for advancing the understanding of multilingual language use, especially in contexts where both Assamese and English are used interchangeably.

Ultimately, this dataset goes beyond just being a collection of sentences - it represents a significant step forward in the study of bilingual communication. Whether for developing new models, testing hypotheses, or expanding our understanding of language processing, it holds the potential to inspire new directions of research and application in the field of computational linguistics.

REFERENCES

[1] Ritchie, William C., and Tej K. Bhatia. "Social and psychological factors in language mixing." The handbook of bilingualism and multilingualism (2012): 375-390.

[2] Bali, Kalika, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook." In Proceedings of the first workshop on computational approaches to code switching, pp. 116-126. 2014.

[3] Younas, Aqsa, Raheela Nasim, Saqib Ali, Guojun Wang, and Fang Qi. "Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches." In 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE), pp. 66-71. IEEE, 2020.

[4] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. "Overview of the track on sentiment analysis for dravidian languages in code-mixed text." In Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 21-24. 2020.

[5] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. "Overview of the track on sentiment analysis for dravidian languages in code-mixed text." In Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 21-24. 2020.

[6] SR, Mithun Kumar, Lov Kumar, and Aruna Malapati. "Sentiment Analysis on Code-Switched Dravidian Languages with Kernel Based Extreme Learning Machines." In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pp. 184-190. 2022.

[7] Srivastava, Abhishek, Kalika Bali, and Monojit Choudhury. "Understanding script-mixing: A case study of Hindi-English bilingual Twitter users." In Proceedings of the 4th Workshop on Computational Approaches to Code Switching, pp. 36-44. 2020.

[8] Pathak, Dhrubajyoti, Sanjib Narzary, Sukumar Nandi, and Bidisha Som. "Part-of-speech tagger for Bodo language using deep learning approach." Natural Language Processing (2024): 1-15.

[9] Raha, Tathagata, Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. "Development of pos tagger for english-bengali code-mixed data." arXiv preprint arXiv:2007.14576 (2020).

[10] Talukdar, Kuwali, and Shikhar Kumar Sarma. "Deep Learning based Part-of-Speech tagging for Assamese using RNN and GRU." Procedia Computer Science 235 (2024): 1707-1712.

[11] Chakravarthi, Bharathi Raja, et al. "Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text." Language Resources and Evaluation 56.3 (2022): 765-806.

[12] Ahmad, G. I., Singla, J., Anis, A., Reshi, A. A., & Salameh, A. A. (2022). Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus: A comprehensive review. International Journal of Advanced Computer Science and Applications, 13(2).

[13] Ahmad, Gazi Imtiyaz, et al. "Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus: A comprehensive review." International Journal of Advanced Computer Science and Applications 13.2 (2022).

[14] Shekhar, Shashi, et al. "Hatred and trolling detection transliteration framework using hierarchical LSTM in code-mixed social media text." Complex & Intelligent Systems 9.3 (2023).

[15] Thara, S., and Prabaharan Poornachandran. "Transformer based language identification for malayalam-english code-mixed text." IEEE Access 9 (2021): 118837-118850.

[16] Hegde, Asha, et al. "Corpus creation for sentiment analysis in code-mixed Tulu text." Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages. 2022.