ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

A Hidden Markov Model Framework for POS Tagging of English–Punjabi Code-Mixed Social Media Text

Sunita[1*], Ajit Kumar[2], Neetika Bansal[3]

[1]Research Scholar, Punjabi University Patiala, Punjab

[2]Associate Professor, Multani Mal Modi College, Patiala, Punjab

[3]Assistant Professor, Punjabi University College of Eng. And Mgt., Rampura Phul

***Corresponding Author**

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150100095

Received: 30 January 2026; Accepted: 07 February 2026; Published: 17 Februar 2026

ABSTRACT

Part-of-Speech (POS) tagging for code-mixed text is notably challenging due to frequent language switching, non-standard orthography, transliteration issues, and the prevalence of informal syntactic structures in user-generated content. This study presents a Hidden Markov Model (HMM)-based approach tailored to English–Punjabi code-mixed text, specifically addressing bilingual interactions in which Punjabi is written in Romanized script. A code-mixed corpus was strategically compiled from platforms such as YouTube, Facebook, and WhatsApp, and meticulously annotated at the token level. The resulting dataset comprises 900 sentences totalling 10,117 words, showcasing diverse mixing patterns and typical social media artefacts, including abbreviations, emojis, and irregular punctuation. The proposed framework conceptualizes POS tagging as a sequence labelling problem. It estimates emission and transition probabilities through the annotated corpus and employs the Viterbi algorithm to decode the most probable tag sequences. Experimental evaluations yield an overall tagging accuracy of 71.52%, establishing a probabilistic baseline for English–Punjabi code-mixed POS tagging. This work lays the groundwork for future research to integrate richer feature sets and leverage neural architectures to enhance performance.

Keywords: code-mixed text, part-of-speech tagging, hidden Markov model (HMM), natural language processing (NLP), language identification

INTRODUCTION

The rapid proliferation of social media has significantly transformed multilingual users' communication practices, resulting in the widespread use of code-mixed language—the integration of two or more languages within a single sentence or discourse [1]. In multilingual countries such as India, code-mixing of English with regional languages, such as Punjabi, has become increasingly common on digital platforms such as Facebook, Twitter, and YouTube. While this linguistic phenomenon is a natural aspect of human communication, it poses challenges for computational language processing systems, which are traditionally trained on monolingual, grammatically standardised text [2].

POS tagging, which entails assigning grammatical categories such as noun, verb, or adjective to words in a text, is a fundamental component of numerous NLP applications, including syntactic parsing, sentiment analysis, and machine translation [3]. In code-mixed contexts, however, POS tagging presents significant challenges due to several factors: irregular grammar, transliteration (typically into Roman script), phonetic spellings, and word-level language switches. These factors frequently lead to data sparsity and reduced accuracy in conventional NLP pipelines that rely on standardised linguistic structures [4].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

In addressing these challenges, statistical sequence modelling approaches, particularly those utilising the HMM, have shown considerable promise in capturing contextual dependencies between words and their associated tags. An HMM is a probabilistic model in which the observable sequence comprises words, and the hidden states correspond to their respective POS tags. By employing statistical estimates of transition probabilities (the relationships between successive tags) and emission probabilities (the relationships between words and tags), the model can deduce the most probable tag sequence for a given text through algorithms such as Viterbi decoding [5].

HMM-based POS tagging has demonstrated considerable success across monolingual and bilingual contexts, primarily due to its effective modelling of sequential dependencies and contextual ambiguity. Nevertheless, its effectiveness in processing English-Punjabi code-mixed data—characterised by mixed scripts, transliterated words, and non-standardised usage—remains largely underexplored. The absence of adequately annotated corpora for code-mixed languages has further impeded the development of robust statistical and deep learning models tailored to these scenarios. This research endeavours to design and evaluate an HMM-based POS tagger specifically developed for English-Punjabi code-mixed social media text. The study utilises a manually annotated code-mixed corpus comprising 10,117 tokens, annotated for both language identification and POS categories in accordance with the Universal POS tagset. The proposed model estimates state transition and emission probabilities using maximum likelihood estimation and employs the Viterbi algorithm to decode the optimal tag sequence.

Experimental evaluations indicate that the HMM-based model balances interpretability and accuracy, achieving notable improvements over unigram and bigram taggers. However, while these statistical models handle contextual relationships well, they still struggle with out-of-vocabulary words, informal transliterations, and sparse data.

The remainder of this paper is structured as follows: Section 2 reviews related work and prior approaches to code-mixed POS tagging. Section 3 outlines the methodologies for dataset creation, preprocessing, and annotation. Section 4 focuses on the HMM framework and the processes involved in model training. Section 5 presents an analysis of experimental results and performance metrics, and Section 6 concludes by discussing potential avenues for future research, including extensions of hybrid and neural models.

LITERATURE SURVEY

Earlier approaches to POS tagging—both monolingual and code-mixed—commonly employed probabilistic sequence models such as HMMs and Conditional Random Fields (CRFs). These models treat tagging as structured prediction over a sequence of tokens, leveraging dependencies between adjacent tags (and, in the case of CRFs, utilising rich token features). In code-mixed scenarios, HMMs serve as strong baselines due to their simplicity and interpretability: emission probabilities capture token–tag associations, while transition probabilities model frequent tag sequences that may occur across languages. However, HMMs exhibit sensitivity to data sparsity and out-of-vocabulary words, which are prevalent in Romanised and informal text.

CRF-based systems generally demonstrate improvements over HMMs by integrating orthographic, contextual, and language-related features, such as character n-grams, affixes, capitalisation patterns, surrounding words, and predicted language ID. Feature-rich CRFs have gained popularity in code-mixed text applications, particularly in shared tasks and early benchmarks, owing to their balance of efficiency and robustness. [1] proposed an intelligent grammatical tagger for a Spanish–Mixtec parallel corpus consisting of 12,300 sentences. They used the EAGLES tagging standard and manually tagged 600 sentences, while semi-automatically tagging 3,000 with GPT-4o. Multiple models, including GPT-4o, CRF, and BERT-based architectures, were trained, with CRF achieving a precision of 0.74 and recall of 0.80.[3] proposed A system for POS Tagging for grammar checking in Punjabi, featuring a tagset that includes Noun, Pronoun, Adjective, Verb, Adverb, Postposition, Conjunction, Interjection, Particles, and Verb-part, totalling around 630 tags. Currently, a rule-based approach for POS tagging in Punjabi achieves nearly 80% accuracy. [4] proposed a POS tagging system for code-mixed English-Hindi

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

text, based on 552 Facebook posts and 554 tweets. Four machine learning algorithms (CRF, Sequential Minimal Optimisation, Naïve Bayes, and Random Forests) were applied using different feature sets.

They utilised a 13 fine-grained and a 39 coarse-grained tagset, achieving an average accuracy of 71.97% on the coarse-grained tagset. [6] A supervised algorithm based on HHM was proposed to tag code-mixed social media posts in Indian languages, including English, Hindi, Bengali, and Telugu. The system was trained and tested using data from Facebook, Twitter, and WhatsApp, focusing on three code-mixed language pairs: Bengali-Hindi, Hindi-English, and Telugu-English. The accuracy of the annotated tags was evaluated using the F-measure, demonstrating the effectiveness of the HMM-based POS Tagger for coarse-grained POS tagging across various language pairs and social media platforms. [7] proposed a POS-tagging system for code-mixed data involving Bengali words in Roman script and English. The approach involved collecting and cleaning English-Bengali tweets. Tokens were tagged based on their language and processed by separate POS taggers for English and Bengali, which then combined their tags into a universal set.

The system achieved an accuracy of 75.29% on manually tagged 5,148 mixed sentences. [8] proposed a deep learning-based approach for Malayalam Twitter POS tagging, comparing RNN and its variants (LSTM, GRU, BLSTM) at both word and character levels. A tag set with 17 coarse tags and 9,915 manually tagged tweets was used for testing. The experiments revealed that the GRU model with 64 hidden states achieved the highest f1-measure of 0.9254 at the word level, while the BLSTM model reached 0.8739 at the character level. Ultimately, the system achieved 84.40% accuracy on the Hindi-English task. [9] proposed a POS tagger specifically for Malay tweets, creating new tags that reflect informal social media language. The researchers used a deep learning approach with a BiLSTM-CRF classifier. Their method combined embeddings from both a Malay Wiki Word2Vec and a Malay Twitter corpus. The BiLSTM-CRF model outperformed traditional classifiers, achieving an F1-Score of 94%. [11] the research focuses on three Indian languages—Hindi, Bengali, and Telugu—mixed with English. It employs machine learning techniques, particularly word2vec for feature extraction, and experiments with Log-linear models for tagging. They used the Stanford POS tagger's machine learning algorithms, including Decision Trees and Naive Bayes. The constrained submission method achieved an accuracy of 75.46%. [12] proposed a POS-tagging system for Urdu tweets, including a corpus of POS-tagged data. They addressed the challenge of limited manually annotated data by exploring bootstrapping and developing a supervised POS tagger. Their experiments involved two pre-trained Urdu taggers tested on well-edited Urdu text and tweets. The tagset includes 33 POS tags.

They manually tagged 500 Urdu tweets and ensured reliability through five-fold cross-validation. Their prepared model for the Stanford POS tagger achieved 93.8% precision, 92.9% recall, and an f-measure of 93.3%. [14] proposed an ensemble POS tagger for Assamese, a morphologically rich language, using a BiLSTM CRF architecture with a 404k token corpus. The best deep learning models achieved an F1 score of 0.746, lower than expected. To improve results, they created a rule-based tagger that utilized Assamese linguistic features, achieving an F1 score of 0.85. By combining the top DL taggers with the rule-based approach, they enhanced the ensemble tagger to an F1 score of 0.925. [15] proposed a trigram HMM-based POS tagger for Indian languages from scratch, using a second-order HHM. The POS-tagged data includes 895 sentences for Bengali (26 tags), 539 for Hindi (25 tags), 1196 for Marathi (27 tags), and 994 for Telugu (24 tags). [16] proposed a supervised POS tagger for Greek social media text was proposed, tackling challenges in NLP for unstructured microblogging.

The authors created the first annotated dataset of 2,405 tweets and a tag set of 22 tags, including Twitter-specific tags. Using Naive Bayes and ID3 algorithms, they achieved an impressive accuracy of 99.87%. [17] proposed a rule-based tokenizer for Portuguese, along with customized tagging strategies based on Universal Dependencies. The DANTEStocks dataset includes 2,737 annotated tweets with tokenization and POS tagging. Tokens are manually tagged with one of 17 possible POS labels. The approach achieved a 98% F1 score for tokenization and 95% for POS tagging, evaluated using accuracy and F1 metrics. [18] A stochastic POS tagger for the Sinhala language was proposed using a HMM with bigram probabilities for training and tagging. The tagger handles multiple POS tags for lexical items and predicts tags for unfamiliar words, utilising the Viterbi algorithm to select the best tag. Social media data served as the annotated corpus, using a balanced set of 24 tags

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

primarily from the UCSC tagset. The tagger achieved an overall accuracy of 63% and nearly 90% for known words, outperforming SVM and hybrid techniques. [19] developed a multi-level annotated corpus of Hindi-English code-mixed text from Facebook, focusing on language identification, normalization, transliteration, and POS tagging. The study used a dataset of 6,983 posts and 113,578 words, achieving 84.6% accuracy in language identification and 79.02% overall POS tagging accuracy. [20] proposed the objective to enhance the accuracy of the existing Punjabi POS tagger, which struggles with resolving ambiguities in compound and complex sentences. To address the POS tagging problem, a Bi-gram HMM was employed. An annotated corpus containing 20,000 words was utilised for training and estimating the HMM parameters. The proposed tagset includes approximately 630 tags, covering various word classes, word-specific tags, and punctuation. During the tagging process, 503 out of the 630 proposed tags were identified in a corpus of 8 million words. The module was tested on a separate corpus consisting of 26,479 words, achieving an accuracy rate of 90.11% as evaluated through a manual approach

While code-mixed POS tagging has been widely studied across various Indian language–English pairs, research specifically focusing on English–Punjabi text in Roman script is limited. Romanisation results in significant lexical variation, creating challenges for handling out-of-vocabulary (OOV) terms and for emission estimation. This highlights the need to investigate an HMM-based POS tagger as a clear, interpretable baseline for Romanised Punjabi code-mixed text. HMM tagging is a well-established classical method for sequence labelling, particularly when enhanced by smoothing and strategies for handling unknown words, providing a solid foundation for comparison with more advanced approaches such as CRFs and transformer-based systems.

METHODOLOGY

Source of data

A new corpus was developed to support POS tagging of English–Punjabi code-mixed text, sourced from platforms where bilingual informal communication is frequent. The raw text was collected from (i) public comments on YouTube related to Punjabi music, entertainment, news, and vlogs, (ii) public posts and comment threads on Facebook, and (iii) WhatsApp chat messages. The corpus contains Punjabi written in Romanised Punjabi (which uses the Latin script) alongside English. The inclusion of Romanised Punjabi is significant given its prevalence in mobile communication and its role in shaping spelling variability in code-mixed contexts.

Dataset for corpus creation

We used an English-Punjabi code-mixed 900 sentences collected from YouTube, Facebook, and WhatsApp groups.

Preprocessing

Preprocessing prepares the dataset for machine learning [8]. Pre-processing improves the effectiveness of machine learning while reducing its training time. This involves data cleaning, tokenisation, and language identification [7]. In our proposed research, in the data cleaning stage, after data collection, we removed duplicate data, translated Punjabi Gurumukhi text to Punjabi Romanised text, removed Hindi Devnagari text, converted emoticons to text, removed extra spaces, reduced continuous punctuation marks, created space between text and punctuation, and removed URLs and hyperlinks. After that, we perform tokenisation, which breaks long text strings into linguistic units, or tokens. An important preprocessing step is the identification of English-Punjabi code-mixed text [6]. We used a dataset of 900 sentences containing 10,117 manually labelled words to train the CRF suite trainer. The model achieved an accuracy of 99.31%.

Tagset defined

There are 16 tags that we use for English and Punjabi code-mixed language such as noun (NOUN), proper noun (PROPN), adjective (ADJ), verb (VERB), auxiliary verb (AUX), adverb (ADV), preposition and postposition

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

(ADP), conjunction (CONJ), interjection (INTJ), particle (PART), negative particle (PART_NEG), cardinal and ordinal numbers (NUM), punctuation mark (PUNCT) and others (X). These tags are used to annotate our corpus.

Manual annotation

Correctly annotating the English-Punjabi code-mixed text is a basic requirement for evaluating the POS tagger's output. Since a new POS tagset is explicitly designed for English-Romanized Punjabi code-mixed social media text. Because no such corpus exists, highlighting the need for manually developing a correctly annotated POS-tagged English-Punjabi code-mixed corpus. To develop the initial corpus, manual annotation of POS tags was performed on randomly sampled 900 sentences containing 10,117 English-Punjabi code-mixed tokens.

Tagged corpus and corpus statistics

After following the corpus development process, we obtain a corpus of 900 sentences containing 10,117 English-Punjabi code-mixed social media tags. The word-level statistics which are given in Table 1 reveal the characteristics of the annotated data set.

Table 1. Word-level statistics

Tags	Count	Percentage Count
ADJ	615	6.08
ADP	927	9.16
ADV	445	4.40
AUX	432	4.27
CONJ	279	2.76
DET	165	1.63
INTJ	47	0.46
NOUN	1935	19.13
NUM	97	0.96
PART	254	2.51
PART_NEG	169	1.67
PRON	789	7.80
PROPN	748	7.39
PUNCT	1283	12.68
VERB	1606	15.87
X	326	3.22
TOTAL	10,117	100

Developed POS tagger for English-Punjabi code-mixed social media corpus using HMM:

The HMM is a statistical sequence model that assigns the most likely sequence of POS tags to a sequence of observed words. In POS tagging, the words are observed symbols, and the tags are hidden states. In the context of code-mixed text, especially English-Punjabi, the challenge increases due to: language switching, informal spelling variations, and the lack of consistent grammar rules across languages.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

Components of HMM

HMM-based POS tagging involves three key probabilities:

- **Transition probability ($P(t_i \mid t_{i-1})$):** Probability of a tag following another tag.
- **Emission probability ($P(w_i \mid t_i)$)**: Probability of a word given a tag.
- **Initial probability ($P(t_1)$):** Probability of a tag starting a sentence.

Training phase

The model is trained on annotated English-Punjabi code-mixed text, where each word is labelled with a POS tag. From this:

- **Transition probabilities** are computed based on tag sequences.
- **Emission probabilities** are computed based on how often a word appears with a given tag.
- **Initial tag probabilities** are derived from sentence-start tags.

For example:

- **Transition:** P (NOUN | PRON) = count (PRON → NOUN) / count (PRON)
- **Emission:** P ("kar" | VERB) = count ("kar" with VERB) / count (VERB)

Viterbi Algorithm

Given a new sentence, the model uses the Viterbi algorithm to find the most probable sequence of POS tags: It builds a probability matrix dynamically. Viterbi decoding (used at the testing phase) accepts a sentence from code-mixed social media text and finds the most likely tag sequence for the test sentence. At each step, it considers all possible previous tags to find the path with the highest probability.

Example:

"I kar homework kal"

Possible tags:

- *I* (PRON), *kar* (VERB), *homework* (NOUN), *kal* (ADV)

The Viterbi algorithm finds the best tag sequence considering both Language context (code-mixing) and Grammar constraints from the training data.

EXPERIMENTS AND RESULTS

For this study, we constructed a corpus of English-Punjabi code-mixed text primarily sourced from Facebook, YouTube, and WhatsApp. The dataset comprises 9,00 sentences, totalling approximately 10,117 tokens. Each token was manually annotated with its corresponding POS tags defined earlier. We implemented the HMM for POS tagging. The HMM assumes that the current POS tag depends only on the previous tag (the Markov assumption), and that each word is generated given the current tag. We used an 80-20 split for training and testing. For training, 720 sentences were used; for testing, 180 sentences.

We evaluated the performance of the HMM-based POS tagger using the following metrics:

- **Accuracy**: Percentage of correctly predicted tags over all tokens.
- **Precision, Recall, and F1-Score**: Calculated for each tag category to assess class-wise performance.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

The HMM model achieved an overall tagging accuracy of 71.52% on the test set. Table 2 provides detailed class-wise precision, recall, and F1 Scores for the selected tags.

Table 2: POS Tagging Performance by Tag Class (Precision/Recall/F1(per tag +macro weighted)

POS Tag	Precision	Recall	F1-Score
ADJ	0.8193	0.5440	0.6538
ADP	0.7024	0.9144	0.7945
ADV	0.7347	0.3956	0.5143
AUX	0.6667	0.8642	0.7527
CONJ	0.8039	0.6406	0.7130
DET	0.6333	0.8636	0.7308
INTJ	0.5000	0.7000	0.5833
NOUN	0.7307	0.6328	0.6782
NUM	0.6923	0.3750	0.4865
PART	0.8889	0.6957	0.7805
PART_NEG	0.8529	0.9355	0.8923
PRON	0.8204	0.7784	0.7988
PROPN	0.4574	0.7679	0.5733
PUNCT	0.7911	0.9766	0.8741
VERB	0.7245	0.5944	0.6531
X	0.7097	0.3056	0.4272
Accuracy			0.7091
Macro Avg	0.7205	0.6865	0.6817
Weighted Avg	0.7270	0.7091	0.7022

CONCLUSION

This study used an HMM to provide a POS tagging method for English–Punjabi code-mixed text. The work addressed the practical need for linguistic processing of multilingual casual content regularly seen on Facebook, WhatsApp, and YouTube, where Punjabi is frequently written in Romanised form and English and Punjabi are often combined within the same utterance. A code-mixed corpus reflecting real-world features such as spelling variation, abbreviations, inconsistent capitalisation, emojis, and irregular punctuation was manually annotated with POS labels (with token-level language cues used during analysis) to facilitate training and evaluation. The HMM-based tagger provides a reliable and comprehensible baseline for this task, as evidenced by experimental data. The model assigns plausible tags based on emission probabilities and successfully captures sequential relationships among tags via transition probabilities. This is especially helpful for frequent closed-class categories and common local syntactic patterns. However, known issues with code-mixed social media content affect performance, including variability in Romanised Punjabi, out-of-vocabulary tokens, and ambiguity between closely related categories (e.g., `NOUN` vs. `PROPN`, `VERB` vs. `AUX`, and discourse markers classified as `PART/INTJ`). These results indicate that larger representations are required to address lexical sparsity and highly variable spellings, even though the HMM framework is appropriate for establishing a foundation and methodically assessing error patterns.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

Overall, the study shows that an HMM-based POS tagger can serve as a strong baseline for English–Punjabi code-mixed NLP and provides a useful annotated resource and analysis for future advances. The corpus, approach, and published findings serve as a basis for advancing POS tagging toward more reliable models and for facilitating downstream uses, including information extraction in bilingual contexts, sentiment analysis, and conversational comprehension.

REFERENCES

1. AlGhamdi, F., Molina, G., Diab, M., Solorio, T., Hawwari, A., Soto, V., & Hirschberg, J. (2016, November). Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 98-107).
2. Baig, A., Rahman, M. U., Kazi, H., & Baloch, A. (2020). Developing a pos tagged corpus of urdu tweets. *Computers*, *9*(4), 90. https://doi.org/10.3390/computers9040090
3. Bansal, N., Goyal, V., & Rani, S. (2020). Experimenting language identification for sentiment analysis of english punjabi code mixed social media text. *International Journal of E-Adoption (IJEA)*, *12*(1), 52-62. https://doi.org/10.4018/IJEA.2020010105
4. Gill, M. S., Lehal, G. S., & Joshi, S. S. (2009). Part of speech tagging for grammar checking of punjabi. *The Linguistic Journal*, *4*(1), 6-21.
5. Jamatia, A., Gambäck, B., & Das, A. (2015). Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. Association for Computational Linguistics.
6. Jamatia, A., Das, A., & Gamback, B. (2020). Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora. Journal of Intelligent Systems, 28(3), 399-408. https://doi.org/10.1515/jisys-2017-0440
7. Kumar, S., Kumar, M. A., & Soman, K. P. (2019). Deep learning based part-of-speech tagging for Malayalam Twitter data (Special issue: deep learning techniques for natural language processing). *Journal of Intelligent Systems*, *28*(3), 423-435. https://doi.org/10.1515/jisys-2017-0520
8. Nikiforos, M. N., & Kermanidis, K. L. (2020, May). A supervised part-of-speech tagger for the Greek language of the social web. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3861-3867). https://aclanthology.org/2020.lrec-1.476/
9. Pakray, P., Majumder, G., & Pathak, A. (2018, January). An hmm based pos tagger for pos tagging of code-mixed indian social media text. In *Annual Convention of the Computer Society of India* (pp. 495-504). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-1343-1_41
10. Pathak, D., Nandi, S., & Sarmah, P. (2023). Part-of-speech tagger for assamese using ensembling approach. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(10), 1-22. https://doi.org/10.1145/3617653
11. Paul, A., Purkayastha, B. S., & Sarkar, S. (2015, September). Hidden Markov model based part of speech tagging for Nepali language. In *2015 international symposium on advanced computing and communication (ISACC)* (pp. 149-156). IEEE. DOI: https://doi.org/10.1109/ISACC.2015.7377332
12. Pimpale, P. B., & Patel, R. N. (2016). Experiments with POS tagging code-mixed Indian social media text. *arXiv preprint arXiv:1610.09799*. https://doi.org/10.48550/arXiv.1610.09799
13. Raha, T., Mahata, S., Das, D., & Bandyopadhyay, S. (2019, December). Development of pos tagger for english-bengali code-mixed data. In *Proceedings of the 16th International Conference on Natural Language Processing* (pp. 143-149). https://aclanthology.org/2019.icon-1.17/
14. Saharia, N., Das, D., Sharma, U., & Kalita, J. (2009, August). Part of speech tagger for Assamese text. In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 33-36). DOI 10.4018/IJSE.2018010102
15. Santiago-Benito, H., Cordova-Esparza, D. M., Castro-Sanchez, N. A., Terven, J., Romero-González, J. A., & Garcia-Ramirez, T. (2025). Automatic grammatical tagger for a Spanish–Mixtec parallel corpus. *SoftwareX*, *29*, 101985. https://doi.org/10.1016/j.softx.2024.101985
16. Sarkar, K., & Gayen, V. (2013). A trigram HMM-based POS tagger for Indian languages. In *Proceedings of the international conference on frontiers of intelligent computing: theory and applications (FICTA)* (pp. 205-212). Berlin, Heidelberg: Springer Berlin Heidelberg.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue I, January 2026

17. Sharma, S. K., & Lehal, G. S. (2011, June). Using Hidden Markov Model to improve the accuracy of Punjabi POS tagger. In *2011 IEEE international conference on computer science and automation engineering* (Vol. 2, pp. 697-701). IEEE. DOI: https://doi.org/10.1109/CSAE.2011.5952600
18. Silva, E. H. D., Pardo, T. A. S., Roman, N. T., & Felippo, A. D. (2021). Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. *Anais*.
19. Tiun, S., Ariffin, S. N. A. N., & Chew, Y. D. (2022, June). POS Tagging Model for Malay Tweets Using New POS Tagset and BiLTSM-CRF Approach. In *ALTNLP* (pp. 160-165).
20. Withanage, S. G., & Silva, T. (2020, November). A stochastic part of speech tagger for the sinhala language based on social media data mining. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 137-142). IEEE. DOI: https://doi.org/10.1109/ICTer51097.2020.9325456
21. Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014, October). Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 974-979).