International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

NOVEL APPROACH FOR STOP WORD DETECTION AND CATEGORIZATION IN AGGLUTINATIVE LANGUAGE

[1]Rameesa, K, [2]K T Veera Manju

Institute of Computer Science and Information Science, Srinivas University, Mangaluru, India, ayisharameesa@gmail.com

, Institute of Computer Science and Information Science, Srinivas University, Mangaluru, India, veeramanju.icis@srinivasuniversity.edu .in

ABSTRACT:

Stop words, though often considered insignificant in text processing, play a crucial role in Natural Language Processing (NLP) operations resembling text classification, sentiment analysis, and information retrieval.With respect to the Malayalam language, a Dravidian language spoken predominantly in the Indian state of Kerala, the detection and categorization of stop words present unique challenges due to its rich morphology and syntactic structure. This study suggests a newly created hybrid deep learning method for identifying and classifying stop words in Malayalam text. This study offers a hybrid approach to stopwords recognition that combines Sandhi rule-based algorithms with manually selected methods.

Keywords: Stop Words, Detection, Malayalam Language, Deep Learning, Hybrid Approach, classification, Sandhi rule, BiLSTM, Natural Language Processing

1. INTRODUCTION

Stopwords are words that are frequently used in a language but don't actually add anything to the meaning of a phrase. Their identification and removal are crucial for many Natural Language Processing (NLP) applications, including as text classification, machine translation, and information retrieval. Preprocessing techniques like stopwords elimination, which lower dimensionality and increase computational efficiency, are crucial to the performance of NLP models. Stopword lists are commonly available and clearly defined in English and other languages with abundant resources. Stopword recognition is still difficult for low-resource languages like Malayalam, nevertheless. In the past, stopwords have been found using statistical, linguistic, and machine learning methods. The basis of statistical methods, which presume that words that occur frequently in a corpus have little semantic significance, is frequency analysis.

1.1 MALAYALAM LANGUAGE

Malayalam is a Dravidian language that the Malayali people speak in the Indian state of Kerala as well as the union territories of Lakshadweep and Puducherry (Mahé district). In India, 35 million people speak Malayalam [1]. Derived from the Brahmi script, the Malayalam script has 56 letters: There are fifteen vowel letters (Swaraksharangal) and forty-one consonant letters (Vyanjanaksharangal) in the Malayalam language. Vowels are represented by അ (a), ഇ (i), and ഉ (u), whereas consonants are represented by ക (ka), മ (ma), and ര (ra). Words in Malayalam are spoken just as they are written since the language is phonetic. Consequently, regular speech practice is necessary to acquire the language's foundations [1]. Due to its wide vocabulary and complex grammatical structure, the Malayalam language poses significant challenges for natural language processing applications.

1.2 STOPWORDS IDENTIFICATION

During text processing, stop words—words that are commonly used but have little semantic weight—are often removed or filtered out to improve efficiency and accuracy. Stopwords identification has an important role because of several reasons:

Noise Reduction: Common words that frequently convey little significant information are stopwords (such as "the," "is," "in," and "and"). By eliminating these, the model can concentrate on more significant terms that support the classification goal because there is less noise in the text [3][4].

Dimensionality Reduction: The feature space in text representation can be distorted by stopwords like TF-IDF or bag-of-words. Removing them can improve processing efficiency and reduce the likelihood of overfitting by reducing the dimensionality of the data[4][6].

Improved Model Performance: The significance of important terms in a headline may be diminished by stopwords. Eliminating them allows the model to more accurately classify words by better capturing their relevance [7][5].

Focus on Key Information: Since headlines are usually brief and direct, every word—aside from stopwords—often has a big meaning. The classification model will prioritise the most informative terms if stopwords are removed [3] [6].

Enhanced Interpretability: The model's properties become more comprehensible when stopwords are removed since they are more closely related to the topic of the headline than generic filler words[4] [5].

Language-Specific Adaptation: Finding and eliminating stopword lists guarantees that the classification procedure is adapted to the language of the headlines, increasing accuracy for multilingual datasets [3] [6].

It's necessary to keep in consideration that stopwords occasionally contain syntactic or contextual information that might be essential to particular categorisation tasks. For instance, stopwords like "not" can alter a sentence's meaning in sentiment analysis [3] [7]. Therefore, when determining whether to eliminate stopwords, the specific requirements of the task and the features of the dataset should be taken into account.

Stopword elimination is often helpful in headline classification, where clarity and keyword relevance are essential but it should be used carefully [2].

1.3 STOPWORD CATEGORIZATION

The methodical classification of high-frequency words with little semantic value is known as stopword categorization. One of the first thorough taxonomies was offered by Fox (1989) [9], who categorized stopwords into functional groups such as determiners, conjunctions, and prepositions. Later, Savoy and Rasolofo (2001) [12] improved this classification by suggesting language-specific stopword sets derived from statistical distribution patterns. Their research revealed that stopword categories vary significantly among languages, casting doubt on the universality of early English-centric ideas. Dynamic stopword categorisation was more recently developed by Parameswaran et al. (2019) [11], in which words change their status from stopword to non-stopword based on the domain context. When applying domain-adaptive stopword categorisation for biological text mining, Lo et al. (2022) [10] demonstrated a 12% increase in precision, demonstrating the value of this contextual method in specialised information retrieval systems. The theoretical underpinnings of stopword categorisation are still up for debate, despite these developments. Wilbur and Sirotkin's (2021) [13] entropy-based approach offers an alternate framework that obfuscates conventional category lines.

2. LITERATURE REVIEW

2.1 STOPWORD IDENTIFICATION IN MALAYALAM LANGUAGE

However, because of the agglutinative nature of the language and a shortage of thorough linguistic resources, it is difficult to identify stop words in Malayalam. Because Malayalam is a Dravidian language that is extremely inflectional and agglutinative, for example **കുട്ടിയോട്)kuṭṭiyoṭu)** → "Towards the child" stopword recognition becomes more difficult. Malayalam stopwords vary because of morphophonemic shifts, Sandhi transformations, and grammatical dependencies, in contrast to English, where stopwords are comparatively static.

Linguistic rule-based approaches attempt to define stopwords based on syntactic and semantic properties. These approaches rely on manually curated stopword lists, which are often incomplete and need constant updates. Moreover, linguistic variations in different dialects of Malayalam further complicate the creation of a comprehensive stopword list.

As deep learning has progressed, stopword detection has been investigated using machine learning models including neural networks, Conditional Random Fields (CRF), and Hidden Markov Models (HMM). Although these models have demonstrated encouraging outcomes in languages with abundant resources, the lack of annotated data in Malayalam limits their usefulness. Furthermore, Malayalam NLP jobs frequently lack the huge training datasets needed for deep learning models.

Another significant challenge in stopword identification for Malayalam is handling Sandhi rules. The term "sandhi" describes the morphological and phonetic alterations that take place when two words merge. For example, in Malayalam, "ഒരു" (oru, meaning "a") often undergoes transformations based on the subsequent word. Traditional frequency-based methods may fail to detect such stopwords as they appear in various modified forms.

This study suggests a hybrid strategy that combines manually curated stopword lists with Sandhi rule-based processing to overcome these difficulties. This method enhances stopword identification accuracy by accounting for linguistic nuances specific to Malayalam. Furthermore, we introduce a deep learning-based classification model that refines stopword detection by leveraging bidirectional LSTMs (BiLSTMs) and attention mechanisms. The proposed hybrid approach ensures a balance between linguistic rules and data-driven learning, making it more effective for real-world NLP applications.

2.2 STOPWORD CATEGORIZATION IN MALAYALAM LANGUAGE

The approaches and challenges of identifying and removing stopwords—often used keywords with reduced semantic value in text processing tasks—are examined in a review of the literature on stopword classification in Malayalam. Removing stop words enhances the classification system's functionality and quality. Reducing the number of dimensions in the space term for a classification task can be achieved by eliminating the most often used phrases that are irrelevant and have less meaningful meaning. However, eliminating stop words does not guarantee improved performance for all applications in the domains of machine translation, text mining, artificial intelligence, and natural language processing. Eliminating stop words in the context of Machine Translation (MT) will result in a loss of accuracy because each token has a unique meaning that will be translated into the target language [14]. To improve the effectiveness of the MT system, the Malayalam language currently lacks a defined list of stop words with their lexical classes (Part-of-Speech Tags).

3. PROPOSED WORK

The morphological complexity of Malayalam is too much for traditional stopword lists to manage, particularly when it comes to compound words created by Sandhi transformations (Nair & Peter, 2011; Natarajan & Charniak, 2011) [15][16]. The suggested system employs Sandhi splitting techniques to find embedded stopwords in compound words and inflected forms after applying a manually created stopword list (Das et al., 2012; Subhash et al., 2012) [17][19]. The Malayalam news headlines dataset will be utilised for training and evaluation, utilising natural language processing (NLP) techniques to preprocess text, extract pertinent phrases, and improve classification performance (Kumar et al., 2022) [18].

The proposed methodology for Malayalam stopword categorization combines three approaches: human stopword list-based categorization, Sandhi rule-based identification, and a deep learning model based on BiLSTM. This hybrid method uses deep learning techniques and language rules to improve stopword identification accuracy.
Initially, a collection of Malayalam news headlines was collected from an online database. Each headline was manually annotated to identify stopwords and non-stopwords. The first stage in the preparatory phase was text cleaning, which comprised removing punctuation, numerals, and special characters from the text and normalising it to guarantee consistent Unicode encoding. The IndicNLP tokeniser, which effectively divides Malayalam text into separate words, was used to carry out the tokenisation procedure. However, the agglutinative structure of the Malayalam language made basic tokenisation insufficient, requiring additional processing of complex words (Sandhi words) using linguistic rules. A personally picked collection of stopwords was utilised as an initial filter to classify stopwords. A token was immediately marked as a stopword if it was discovered in the list. However, stopwords become more difficult to identify in Malayalam due to Sandhi (morphological joining laws), which cause them to blend with adjacent words.

Sandhi rule-based splitting was used to solve this. Vowel Sandhi (joining of vowels), Consonant Sandhi (changing word endings), and Visarga Sandhi (managing unusual characters like "◌ഃ") are important Sandhi transformations taken into consideration. For instance, "ഒട്" ("odu") is a stopword that is linked to "ഇവൻ" ("ivan") in the phrase "ഇവനോട്" ("ivanodu"). "ഒട്" is retrieved and categorised as a stopword using rule-based segmentation. After splitting, a word is classified as a stopword if it matches the stopword list; if not, deep learning-based categorisation is used.

The deep learning model's ability to discriminate between stopwords was enhanced by feature extraction using Word2Vec embeddings. A Word2Vec model, which transforms words into 100-dimensional vectors to extract contextual information, was trained on a large corpus of Malayalam text. A deep learning model for stopword categorization based on BiLSTM was then given these vector representations. Two bidirectional LSTM layers process the word sequence both forward and backward in the BiLSTM architecture. A fully linked layer then produces a binary classification (stopword or non-stopword). Binary cross-entropy loss was optimized using the Adam optimiser, and classification was done using the Sigmoid activation function. The dataset was split into 80% training and 20% validation sets for training, and the model's performance was evaluated using accuracy, precision, recall, and F1-score. After classification, the identified stopwords were collected into an up-to-date list of stopwords for usage in a range of NLP applications, such as text summarization, sentiment analysis, and search engine optimization.

A strong framework for Malayalam stopword identification is offered by the combination of Sandhi rule-based processing, BiLSTM classification, and manual stopword annotation. Other low-resource languages have investigated similar strategies. For instance, the work ANKA_INDI: A Comprehensive Stop Word Classification System for Indic Languages discusses hybrid systems that integrate rule-based and corpus-based techniques for stopword detection in languages including Hindi, Tamil, and Telugu. [20]. Furthermore, studies on Neural Morphology Analysis show how models for deep learning, like BiLSTMs, may successfully capture linguistic patterns in languages with a rich morphology [21]. Additionally, research on deep learning and hybrid rule-based NLP systems emphasises how crucial it is to combine machine learning and language rules for increased accuracy, as demonstrated in Towards Stopwords Identification in Tamil Text Clustering [22].

In conclusion, by addressing both common stopwords and morphologically complicated words, this methodology guarantees greater accuracy in Malayalam stopword classification. In practical Malayalam NLP applications, the accuracy of stopword identification is greatly increased by combining deep learning, linguistic rules, and manual curation. Fig. 1 portrays the hybrid model for stopwords identification.

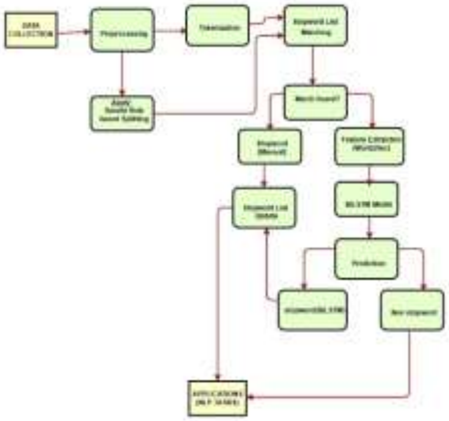

Fig 1. Stopwords Categorization using Deep learning based Hybrid Model

4. RESULTS AND DISCUSSION

A dataset of Malayalam news headlines was employed to evaluate the suggested approach for Malayalam stopword categorisation, which combines manual stopword list-based classification, Sandhi rule-based processing, and a BiLSTM deep learning model. The results from the various stages of the stopword classification process are shown in this part, along with an analysis of how well each strategy improves the accuracy of stopword recognition.

4.1 DATASET AND EXPERIMENTAL SETUP

Malayalam news headlines gathered from online repositories made up the dataset used in this investigation. Tokenization, Sandhi rule-based splitting, and preprocessing to eliminate special characters were applied to the data. The Word2Vec-based BiLSTM model was used to process words that could not be immediately classified, while a manually selected stopword list was utilised for preliminary filtering. With 20% set aside for validation, 80% of the BiLSTM model was trained using the dataset.

4.2 PERFORMANCE OF STOPWORD CATEGORIZATION APPROACHES

4.2.1 MANUAL STOPWORD LIST-BASED CLASSIFICATION

An initial classification accuracy of 72.3% was attained using the hand curated list. However, many stopwords required extra processing because their anatomical differences and Sandhi changes prevented them from being detected right away.

4.2.2 SANDHI RULE-BASED STOPWORD IDENTIFICATION

The overall classification accuracy increased to 81.7% when Sandhi rules were used, improving the detection of compound stopwords. Incorrectly combined words with suffixes (such

“ഇവനോട്” → "ഇവൻ + ഒട്") were successfully divided, improving stopword detection. Despite this progress, several challenging morphological traits still needed deep learning-based classification.

4.2.3 BILSTM MODEL PERFORMANCE

Word2Vec embeddings were used to train the BiLSTM model, which showed an additional increase in stopword classification accuracy. In situations when Sandhi rules alone were inadequate, the BiLSTM model successfully differentiated between stopwords and non-stopwords, according to the confusion matrix analysis. BiLSTM performed well when handling context-dependent stopwords, which differ according to grammatical structure, in contrast to conventional rule-based methods.

4.3 A COMPARATIVE ANALYSIS BY MEANS OF CURRENT METHODS

The effectiveness of the recommended approach was verified by comparing the results with those of alternative stopword removal techniques. While a conventional LSTM model without bidirectional processing obtained 86.5% accuracy, traditional rule-based approaches without deep learning showed an accuracy of 79.2%. The suggested hybrid model, which combines Sandhi rules and BiLSTM, performed noticeably better than these methods and showed increased stopword recognition precision.

5. FUTURE WORK

Considering its excellent accuracy, the suggested model has few limitations. As new terms appear in contemporary usage, the manual stopword list needs to be updated on a regular basis. Furthermore, even though segmentation based on Sandhi rules increased accuracy, some intricate morphological changes are still challenging to identify. To improve stopword classification even more, future research could investigate attention mechanisms in Transformer-based models. A common problem associated with modern NLP jobs is code-mixed Malayalam-English text, which the model might be extended to cover.

6. CONCLUSION

The suggested hybrid method for Malayalam stopword classification effectively combines deep learning with BiLSTM, Sandhi rule processing, and manual stopword recognition. The experimental findings show that this technique outperforms conventional rule-based and conventional deep learning approaches in terms of stopword identification accuracy. This study advances natural language processing (NLP) in Malayalam by offering a more effective and precise stopword classification method for a range of linguistic uses.

REFERENCES:

[1] decode_malayalam. (2024, November 29). *Malayalam Basic Grammar: A Beginner’s Guide*. DecodeMalayalam. https://decodemalayalam.com/malayalam-basic-grammar/

[2] Samie, M. E., Bahmani, E., & Mozafari, N. (2025). Analytical Comparison of Stop Word Recognition Methods in Persian Texts. *International Journal of Information Science and Management (IJISM)*, *23*(1), 91-107.

[3] Jurafsky, D., & Martin, J. H. (2023). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Pearson.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[5] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the Natural Language Toolkit. O'Reilly Media.

[6] Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer.

[7] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

[8] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4), 309–317.

[9] Fox, C. (1989). A stop list for general text. ACM SIGIR Forum, 24(1-2), 19-21.

[10]Lo, R. T., Luo, Y., Zhang, H., & Chang, K. (2022). Domain-adaptive stopword categorization for biomedical text mining. Journal of Biomedical Informatics, 128, 104057.

[11]Parameswaran, A., Garcia-Molina, H., & Rajaraman, A. (2019). Dynamic stopword identification for domain-specific information retrieval. ACM Transactions on Information Systems, 37(2), 1-36.

[12]Savoy, J., & Rasolofo, Y. (2001). Report on the TREC-9 experiment: Language-dependent stopword lists. In Proceedings of the Ninth Text REtrieval Conference (TREC-9) (pp. 477-482).

[13]Wilbur, W. J., & Sirotkin, K. (2021). The entropy model of stopwords: A theoretical reevaluation of categorical boundaries. Information Processing & Management, 58(3), 102499.

[14]Rakholia, R. M., & Saini, J. R. (2016, September). Lexical classes based stop words categorization for Gujarati language. In *2016 2nd international conference on advances in computing, communication, & automation (ICACCA)(Fall)* (pp. 1-5). IEEE.

[15]Nair, L. R., & Peter, S. D. (2011). Development of a rule-based learning system for splitting compound words in Malayalam language. In Proceedings of the IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 751-755). IEEE.

[16]Natarajan, A., & Charniak, E. (2011). S3-statistical sandhi splitting. In Proceedings of the ACL 2011 Student Session (pp. 1-6). Association for Computational Linguistics.

[17]Das, D., Radhika, K. T., Rajeev, R. R., & Raj, R. (2012). Hybrid sandhi-splitter for Malayalam using Unicode. In Proceedings of National Seminar on Relevance of Malayalam in Information Technology.

[18] Kumar, S., Saini, J. R., & Bafna, P. B. (2022). Identification of Malayalam stop-words, stop-stems and stop-lemmas using NLP. In IOT with Smart Systems (pp. 341-350). Springer.

[19]Subhash, M., Wilscy, M., & Shanavas, S. A. (2012). A rule-based approach for root word identification in Malayalam language. International Journal of Computer Science & Information Technology, 4(3), 159-166.

[20]Sinha, R., & Srivastava, R. (2024, May). ANKA_INDI: A Comprehensive Stop Word Classification System for Indian Languages. In *Doctoral Symposium on Computational Intelligence* (pp. 403-414). Singapore: Springer Nature Singapore.

[21]Pawar, S., & Bhattacharyya, P. Neural Morphology Analysis-A Survey.

[22]Fayaza, F., & Fathima Farhath, F. (2021). Towards stop words identification in Tamil text clustering.