

Original Article

Hybrid Contextual Embedding Framework for Low-Resource Gujarati Word Sense Disambiguation

Avani N Dave¹, Sanjay M Shah², Nakul R Dave³

¹Computer/IT Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India.

²Computer Engineering Department, Government Engineering College, Rajkot, Gujarat, India.

³Computer Engineering Department, Vishwakarma Government Engineering College, Ahmedabad, Gujarat, India.

²Corresponding Author : sanjay_shah_r@yahoo.com

Received: 19 May 2026

Revised: 19 August 2026

Accepted: 26 August 2026

Published: 30 September 2026

Abstract - Word sense disambiguation continues to constitute significant challenges in natural language processing, particularly for languages such as Gujarati that exhibit considerable morphological variation. Thus, natural language processing has not progressed well for word-sense disambiguation in Gujarati. After all, there are essentially two issues here: insufficient data and the languages' intrinsic complexity. Given the various forms of Gujarati and the widespread use of idioms, models do not perform well and yield inaccurate results. In order to address a gap in the existing literature, this study presents a new Gujarati word sense disambiguation dataset, which has been manually word-sense-tagged and is specifically tailored for ambiguous Gujarati lexemes. The corpus includes 150 ambiguous words, each of which has been carefully annotated to show its exact meaning in context. This dataset contains 426 distinct meanings across 4,400 sentences, making it a complete test set for sense tagging. Still, natural language processing has a long way to go before it can correctly interpret a written word and universally understand its meaning in highly morphological languages like Gujarati. This paper introduces a context-aware hybrid architecture designed to improve the identification of polysemous words in Gujarat texts. The proposed model utilizes various contextual characteristics and combines lexical, statistical, and contextual components to address the problems associated with the use of single-component models. The testing process involved the use of five-fold cross-validation considering Random Forest and BERT algorithms. The resulting accuracy was remarkably high at 97.54%, with the corresponding Macro-F1 value being 97.40%. A suite of models that exhibits tolerance for the correct sense of the word without overfitting can be found in the well-balanced macro-precision value of 98.14%. The system's stability and consistency have been confirmed by the consistently low standard deviation, which is ± 0.53 percentage points. Together, these results show that the method is viable when applied to word sense disambiguation in the Gujarati language.

Keywords - Machine learning, Natural Language Processing, Word Sense Disambiguation, BERT, Random Forest, Sense Annotated Corpus, Gujarati Language, Polysemy Words.

1. Introduction

Word Sense Disambiguation (WSD) is an important computing task associated with the determination of the specific meaning of a word with multiple senses based on its context in Natural Language Processing (NLP) [1-3]. Ambiguity arises in language because of the context-based nature of lexical analysis. Words like "bank" can refer to financial institutions as well as riverbanks.

The analysis of the linguistic ambiguities present in languages spoken in India has increasingly resorted to more advanced models based on transformers and hybrid systems because of their capability to model semantic distinctions that cannot be achieved with the help of conventional lexicons. Such sophisticated models employ self-attention functions and contextualized word embedding models like BERT to

incorporate intricate syntactic and morphological relationships not captured by early approaches [4-6]. Due to the statistical, lexical, and contextual features, there is a fruitful way to disambiguate polysemy in both local and global discourse contexts.

Therefore, advances in technology are important to handle morphologically rich Indian scripts; in the long run, this will result in high-resource alignment and simplify translation.

These subtle variations through a sentence can fundamentally change its meaning, leading to WSD being called an "AI-complete" problem and an intermediary step for high-end applications of this system, such as machine translation, sentiment analysis, and information retrieval. One



more reason WSD and Gujarati Computational Linguistics have yet to evolve is that it is a morphologically rich, resource-scarce language with a user base of more than 55 million worldwide. Alternatively, at least, people developing WSD for Gujarati are in the early stages.

The complexity of WSD is further accentuated by several structural peculiarities, particularly subtle differences in meaning and idiomatic phrasing, and it is best illustrated by lexical ambiguity in the Gujarati script as shown in Table 1.

Table 1. Example of Lexical Ambiguity for the Gujarati words

Word	Meaning	Gujarati Sentence	Transliteration	English Translation
'વર્ગ' (Varg)	Class/Category	વિદ્યાર્થીઓ નવા વર્ગમાં ગયા.	Vidhyārthiō navā vargmām gayā.	The students went to the new class.
	Square (Math)	ચારનો વર્ગ સોળ થાય છે.	Chārno varg sōḷ thāya chē.	The square of four is sixteen.
'રસ' (Ras)	Juice	મને કેરીનો રસ ભાવે છે.	Manē kērīnō ras bhāvē chē.	I like mango juice.
	Interest	તેને સંગીતમાં ઘણો રસ છે.	Tēnē saṅgītmām ghaṇō ras chē.	He is very interested in music.

Looking for sense in words is a fundamental problem in NLP because the centres of meaning are highly complex. In Gujarati, this is primarily because, alongside its morphological richness, data sparsity in characteristics is found in low-resource language settings. Multiple lexemes-in this case, "વર્ગ" (Varg) and "રસ" (Ras)-require sophisticated types of analyses for determination, such as whether academic meaning, mathematical meaning, or abstract meaning. Single-approach systems tend to stall because they rely on a single set of strategies to automate tasks.

Most of the time, these approaches perform poorly on paper and yield low accuracy. With a top-down approach, high-precision prediction requires a pathway that bridges the accuracy gap between lexical and contextual configurations by leveraging strong statistical features within the software system itself to deliver good practical results. This leads to the prospect of making a further advance beyond basic dictionary methods in place and achieving the precision required for regional language processing.

1.1. Motivation

The fact that regional Indian languages are not being properly handled at the universal level of NLP shows that this study is relevant. Although well-resourced languages benefit from large datasets and sophisticated models, Gujarati is still considerably neglected because of the "knowledge acquisition bottleneck" and the persistent lack of data. As a result, there is a significant research gap concerning the limitations of the existing models; the current baseline systems are unable to deal accurately with the inherent morphological complexity and high level of ambiguity found in the Gujarati language when used in such data-scarce environments. Therefore, it is the viewpoint that this study will bridge the transition taking place in the domain of technologies that were initially dependent on dictionaries to those related to hybrid methods, where context is considered with respect to the structural nuances of the Gujarati script. The development of the "Guj-Sense corpus" is an attempt to establish baselines for the application of well-established machine learning models that will provide precision in distinguishing various polysemous

idioms in their contextual use.

While there have been advancements in WSD for resource-rich languages, the use of WSD for Gujarati is still difficult due to there being a lack of sense-annotated corpora and required computational resources in Gujarati. Gujarati exhibits morphological variations and context-dependent lexical ambiguity, where the intended sense of a word may change according to its surrounding words and sentence context. Existing Gujarati WSD studies are inadequate, as only a few studies have been carried out.

All the studies conducted so far have been restricted to knowledge-based, rule-based, or feature-based approaches. Issues with the previous work done in terms of Gujarati WSD arise out of a lack of datasets, which have been sense-annotated. The results here highlight the importance of the establishment of a framework of WSD for the Gujarati language, which takes into account contextual semantics along with the linguistic information as well.

In addition, the research aims to set a new benchmark in terms of the accuracy standard in future regional language processing. By combining BERT-based contextual embeddings with statistical co-occurrence features into a context-aware hybrid framework, the research aims to handle the performance gap in current single-model-based systems. With these, the ultimate motivation is to equip regional language speakers with efficient tools for machine translation and sentiment analysis that are as reliable as those of high-resource languages.

1.2. Research Questions

This research aims to answer the following key research questions:

1. How significantly does the development of a carefully annotated "Guj-Sense Corpus" alleviate the issues related to data sparsity and enhance model reliability?
2. How can a context-aware hybrid system be developed to accurately identify the intended sense of ambiguous

words in a morphologically rich language such as Gujarati?

3. Can integrating Transformer embeddings, such as BERT, co-occurrence data, and lexical attributes (like Lemma ID) lead to improved WSD results compared to any single traditional method by itself?
4. How reliable and consistent performance can be obtained when utilizing a Random Forest ensemble model with 5-fold cross-validation for WSD in a regional language?

1.3. Contribution

The main contributions of this paper are threefold and are visualized in Fig.

1. **Guj-Sense Corpus:** This paper describes the development of the Guj-Sense corpus, which is a hand-crafted dataset consisting of sentences translated into the Gujarati language only. The corpus comprises 150 different spoken words and includes 426 senses spread across

4,400 sentences. This corpus provides a reliable benchmark to cope with the lack of annotated corpora for Gujarati NLP.

2. **Context-Aware Hybrid Framework:** This research provides an extensive review of the proposed hybrid framework, which combines lexical, statistical, and contextual approaches. The use of character combinations to merge diverse knowledge aspects demonstrates superior performance compared to individual computational methods in addressing the complexities of natural language.
3. **High-Performance WSD Benchmark:** The study proposes a high-performance benchmark by combining the Random Forest and BERT models. The resulting system is highly accurate, showing its high reliability and robustness for identifying polysemous Gujarati texts.

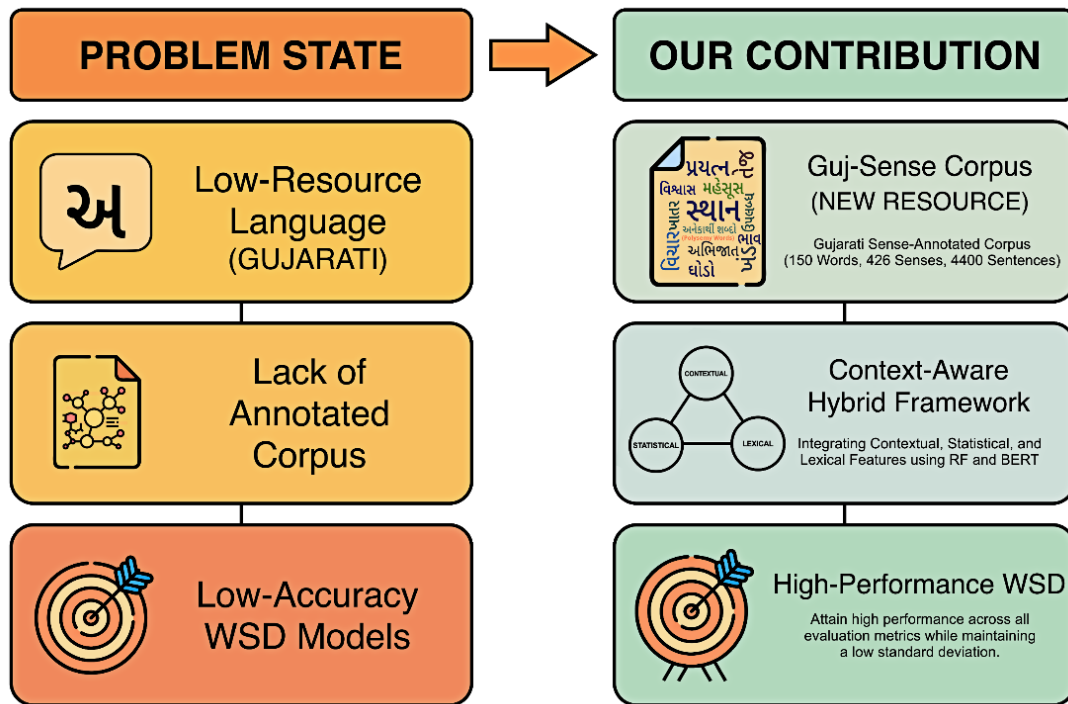


Fig. 1 Contribution of the research paper

1.4. Paper Structure

The remainder of this paper is organized as follows. Section 2 presents a review of related work, focusing on WSD methods relevant to morphologically rich and low-resourced languages. Section 3 details the proposed methodology and describes the architectural design of the context-aware hybrid framework. The experimental setup, the development of a sense-annotated Guj-Sense corpus, performance benchmarks, evaluation metrics, and an exhaustive discussion of the results of the model are presented in Section 4. Section 5 concludes with a summary of the principal findings and potential future research directions.

2. Related Work in Word Sense Disambiguation

This section presents the background studies and literature reviews related to the current work, thereby justifying the need for a new methodology. The history of WSD has been explored, with the focus of research shifting from classic rule-based models and moving through modern machine learning and hybrid systems to the most recent deep contextual neural architectures. WSD is an elementary problem in NLP, which aims to state different meanings of the polysemy word in an automatic fashion, in the context in which it is used. It is considered an AI-complete problem since semantic interpretation demands linguistic, contextual, and

world knowledge [7, 8]. WSD is important in applications like machine translation, information retrieval, question answering, and sentiment analysis [3, 9]. Over time, research on WSD progressed from rule-based to statistical learning, and lately, deep contextual neural networks.

2.1. WSD in High-Resource vs. Low-Resource Languages

Remarkable success has been achieved in WSD for high-resource languages, particularly English, owing to large annotated corpora and the presence of lexical resources such as WordNet. Early approaches relied on supervised learning methods using hand-crafted linguistic features and probabilistic classifiers [2, 10]. Later, distributional semantics and word embedding models improved semantic representation by learning contextual relationships among words [11]. The introduction of deep contextual embeddings has been a breakthrough in the development of models that adapt word representations symbiotically with their context.

WSD's success in English has not been transferred effectively to low-resource languages. Indian languages are characterized by complex morphology, free word order, compounding, and contextually determined meaning. These properties lead to a high degree of data sparsity that significantly interferes with the effectiveness of purely statistical methods. This leads to the problem of severe data paucity, effectuating the inefficacy of the simple statistical means. These studies highlight a lack of annotated resources and lexical data for low-resource languages, which restricts the supervised training of models [12, 13]. Consequently, WSD accuracy for regional languages remains significantly lower than in English.

Low-resourced languages gradually suffer from morphological richness due to inflectional and suffixal morphology. This leads to confusion of senses by context models. Languages also exhibit morphological richness, which alters word forms through inflectional and suffixal morphology, leading contextual models to confuse senses [14]. In several cases, various semantic categories are represented by the same lexical form, which depends on the syntax and domain context. Therefore, the successful ways used in high-resource settings may not be directly applicable to languages such as Gujarati without linguistic adaptation.

2.2. WSD Research in the Gujarati Language

The major reason is the lacunae of large sense-annotated corpora and standard benchmarks. Gujarati WSD is still in an early stage compared to Hindi and English. One of the primary reasons is the lack of large, sense-annotated corpora and standardized benchmarks. The Gujarati language exhibits extensive inflectional variations, dialectal diversity, and semantic ambiguity, which complicate automatic interpretation [15]. Polysemous words have various interpretations, which makes computational disambiguation challenging.

Researchers in early studies of Gujarati WSD used knowledge-based methods that relied on the overlap of dictionary definitions. The authors developed Lesk-based algorithms, which demonstrated that knowledge-based approaches could identify the correct meanings of Gujarati words through contextual analysis of WordNet glosses [16]. Other authors showed that Gujarati WordNet-based methods that determine which word meaning to use by comparing sentence context with lexical semantic relations, produce limited performance because it depended on sparse lexical resources [17]. The survey of Gujarati WSD approaches demonstrated that three different method types, which include knowledge-based methods, supervised techniques and unsupervised methods, have been studied, yet none achieved robust accuracy because contextual variations present modeling challenges. Researchers conducted further research to investigate both heuristic methods and optimization-based techniques.

The authors developed a genetic algorithm-based WSD system that uses lexical knowledge to solve disambiguation through classification search methods [18]. The methods used in this research depend on two main components, which are custom-built rules and WordNet coverage, limiting scalability. After these initial attempts, researchers shifted their attention to learning algorithms within the framework of supervised learning. Vyas and Ganatra proved through supervised classification that machine learning algorithms surpass rule-based algorithms despite depending heavily on labeled data [19]. Also, supervised classification algorithms like decision tree models and Random Forest (RF), with the help of sense-labeled datasets, gradually improved their accuracy. However, the NLP research for the Gujarati language suffers from serious issues because of the unavailability of labelled data and complex morphological structure analysis.

There are three main problems that need to be addressed in WSD, namely, issues of polysemy and homonymy, contextuality, and sense knowledge. Polysemy is described as a case where a particular word has different meanings that are somehow related to each other, while homonymy refers to cases where two words have exactly the same form but different meanings that bear no relation whatsoever. Take, for example, the word “પાન” (Pān), which can mean both a leaf or a page. Here are some examples to show the same: Instance A: “ઝાડ પર લીલું પાન છે.” (Jhāḍ par līlūṁ pān chē. / “There is a green leaf on the tree.”). Instance B: “આ ચોપડીનું પાન ફાટી ગયું.” (Ā chōḍāḍinūṁ pān phāṭī gayūṁ. / “The page of this book tore.”). In Instance A, the adjacent words “ઝાડ” (tree) and “લીલું” (green) give vital context to the NLP model in deriving the interpretation of meaning in the botanical context. In Instance B, the terms “ચોપડી” (book) and “ફાટી” (tore) direct the interpretation toward the meaning of a sheet of paper.

Context identification is only one part of the process. A system also needs sense knowledge, which is an organised inventory that maps all possible interpretations a word may have. To illustrate the extent of lexical ambiguity in the language, Fig shows a word cloud highlighting the most prominent અનૈકાર્થી શબ્દો (Anēkārthī śabdō / polysemous words) found in Gujarati text.



Fig. 2 Word cloud of અનૈકાર્થી શબ્દો (Anēkārthī śabdō / Polysemy Words) for the Gujarati language

Researchers require comprehensive lexical databases to implement machine learning systems that use this sense-knowledge. The Gujarati WordNet provides a valuable resource for researchers studying regional Indian languages. Within this paradigm, meanings of words are organized into distinct synsets, which are groups of synonyms. Each synset has a corresponding definitional gloss, which acts as a common classification label for WSD systems.

2.3. WSD Advancement in Other Indian Languages

Research on regional WSD in Indian languages has progressed through comparative studies across various languages. For Hindi WSD, previous research work used IndoWordNet alongside supervised learning algorithms [20]. For Assamese WSD, researchers use supervised learning algorithms to find patterns in context [21, 22].

Several pieces of research highlight the superiority of machine learning over rule-based approaches when the former has access to annotated corpus data [23]. Studies on Hindi WSD indicate that linguistic variation, compound formation, and morphology are the major hindrances to achieving

disambiguation [12]. Hybrid methods, both supervised and unsupervised, have been explored as well. The use of ensembles of classifiers has shown promising results [13].

In this regard, an alternative approach has been the use of semi-supervised techniques that are not completely dependent on labelled datasets [24, 25]. However, data unavailability and linguistic complexities are still issues even with more efforts in improving performance in Indian languages.

2.4. Transition Towards Neural and Transformer-Based Models

To overcome the drawbacks of symbolic methods, the idea of distributed representations and word embeddings was developed among researchers. WSD methods have moved from the traditional rule-based and statistics-based approaches to more complex neural and contextual learning algorithms. In early machine learning, feature engineering was done manually, and classifiers such as Support Vector Machines, Naïve Bayes, and Decision Trees were used to classify word meanings.

These models involved feature engineering and annotated datasets and produced mediocre results [26-30]. Distributed representations and word embeddings were proposed to overcome the limitations associated with symbolic representations.

The following innovations improved performance in measuring semantic similarity among words compared to rule-based models [4, 11, 31]. With the development of recurrent neural networks and deep learning models, algorithms were developed to automatically identify word-context associations, thus enhancing semantic understanding [26, 27]. However, early neural models suffered from data scarcity, limited vocabulary usage and limited access to essential data [32, 33].

The development of Transformer architecture is a notable improvement for WSD. Context-based language models like BERT and similar models produce dynamic word representations through bidirectional analysis and the ability to recognise distant word relationships, making them more effective for word sense prediction than static word embeddings [34-36].

The relationships between words are inferred from contextual information provided by neighbouring words, allowing the model to estimate the probability of multiple possible senses. Multilingual versions of such models, such as Multilingual BERT and IndicBERT, among others, have been effective for cross-lingual transfer learning and for handling languages with complex morphological structures [37, 38].

Contextual methods analyze regional text corpora to uncover hidden semantic connections among words, thereby

enhancing understanding across multiple linguistic domains [39]. Using attention-based contextual embedding frameworks that integrate hierarchical and local-global attention mechanisms leads to greater disambiguation accuracy for complex sentence structures.

The transformer architectures still face persistent problems when applied to languages with complex morphology and scarce data. Contextual embedding, however, is good at recognising the semantic relationship between words, but it often fails to identify exact word matches, morphological variants, and frequent usage patterns.

The scarcity of data and the uneven distribution of data lead to errors in the classification of infrequent word senses and idioms [14, 40]. So it can be concluded that if we only use a single deep learning model, the efficiency in WSD cannot be guaranteed.

Recently, it has become evident that overcoming these limitations requires the combination of different information sources. For instance, hybrid approaches to WSD that use both contextual embeddings, statistical co-occurrence knowledge, and lexical knowledge can accurately identify meanings better than the use of each type separately [3, 41-43]. RF models, which use ensemble learning, achieve greater robustness by modelling complex non-linear associations across different types of feature sets used to distinguish sense [44-47].

Furthermore, the combination of hybrid learning algorithms with attention-based systems boosts the results in under-resourced scenarios, especially when deep contextual algorithms are combined with traditional learning methods [6, 48]. Newly conducted research in the area of WSD shows that effective feature fusion requires merging contextual, statistical, and lexical features.

3. Design of the Gujarati Word Sense Disambiguation using Context-Aware Hybrid Framework

Design and mathematical structure of the proposed context-aware WSD system for the Gujarati language will be discussed in the following section. Through a well-defined notation framework, the process of disambiguation deals with the issue of word sense ambiguity through feature extraction in stages, beginning with raw text data.

These core algorithmic processes include attention-based word embedding extraction using multilingual BERT, followed by statistical co-occurrence vector formation and finally lexical mapping. The unprocessed text data is subjected to mathematical transformations that generate complex, high-dimensional feature vectors, which are then used by the overall RF ensemble model for classification tasks.

3.1. Problem Formulation and Mathematical Notations

The process of identifying polysemous words can be described in detail with the help of the dataset D , which consists of N samples that can be analyzed as shown in equation 1:

$$D = \{(s_i, t_i, z_i)\}_{i=1}^N \quad (1)$$

Within this set, each instance contains three primary components:

- s_i denotes the input sentence structured as a string.
- t_i represents the target lemma (the ambiguous word), such as the Gujarati term "વડા".
- z_i indicates the true sense label associated with the target word in that specific context, such as "શ્રેણી" (category) or "ગણિત" (mathematics).

The entire vocabulary of target lemmas is defined as $L = \{l_1, \dots, l_L\}$. For any given lemma l , a specific set of possible senses is maintained, denoted as $S(l) = \{s_1^{(l)}, \dots, s_{m_l}^{(l)}\}$, where m_l represents the total number of distinct meanings for that lemma.

3.2. Algorithmic Workflow of Hybrid Framework for Gujarati Word Sense Disambiguation: Integrating Contextual, Statistical and Lexical Features

The hybrid architecture performs its tasks using a process-oriented approach in which data is processed in several phases, along with combining various kinds of features. The computation is done using an unstructured Gujarati text in a sequential process, which necessitates full text pre-processing prior to moving on to higher-level extractions. To describe the computation process in mathematical form, Table below shows the notations. These unambiguous notations define the underlying parameters of the system, as far as they provide the means to form the exact relationship among data instances D , the contextual sentences s_i , the ambiguous target lemmas t_i , and their corresponding true sense labels z_i . Algorithm 1 systematically details the execution steps of the context-aware hybrid Gujarati WSD framework, the comprehensive structure of which is outlined in Table.

The entire life cycle of the disambiguation task is contained within the algorithm: the process begins with the ingestion of data together with the development of semantic co-occurrence dictionaries, and it continues through the extraction process of hidden states from transformer models and is turned with a global RF ensemble for final sense-label prediction and computation of evaluation metrics.

Table 2. List of Notations

Notation	Identifier (Description)
D	The complete sense-annotated input dataset.
S	The set of all sentences extracted from dataset D.
T	The set of all unique target words in dataset D.
s_i	The i^{th} sentence in dataset D.
t_i	The target ambiguous word corresponding to sentence s_i .
y_i	The gold sense label corresponding to t_i .
CoocDict	The co-occurrence sense dictionary stores sense-index mappings for each target word.
Y_t	The set of all possible senses associated with the target word t .
LE_t	Global Label Encoder for target words.
LE_s	Global Label Encoder for sense labels.
c	Total number of unique senses across the entire dataset.
t_i^{enc}	Encoded numeric representation of the target word t_i .
y_i^{enc}	Encoded numeric representation of sense label y_i .
X	Complete feature matrix containing all X_i .
e_i	Contextual CLS embedding for sentence s_i .
$h_j^{(L)}$	Hidden representation of token j at BERT layer L .
L	Total number of transformer layers in BERT.
r	Summation index used to iterate over the last four transformer layers in BERT while computing the sentence embedding.
d	Hidden dimension size of BERT embeddings.
I_i	Set of token indices corresponding to the target word t_i .
W_j	Tokenized sentences.
t_i^{bert}	Contextual embedding of the target word
X_i	Final combined feature vector for instance i .
y	Global encoded sense label vector.
Z_i	Composite label combining encoded target word and encoded sense, for instance, i .

Table 3. Algorithm for Hybrid framework for Gujarati word Sense Disambiguation: Integrating Contextual, Statistical, and Lexical Features

Input:

- Dataset D containing columns — Sentence, Target_word, Sense

Output:

- Trained a global Random Forest model RF
- Target Word LabelEncoder LE_t
- Sense LabelEncoder LE_s
- Evaluation Matrix

Begin

1. Load dataset D.

2. Normalize all sentences in D using Unicode normalization and whitespace cleaning.

3. Extract the set of all sentences (Equation 2):

$$S = \{s_i | (s_i, t_i, y_i) \in D\} \quad (2)$$

4. Extract the set of unique target words (Equation 3):

$$T = \{t_i | (s_i, t_i, y_i) \in D\} \quad (3)$$

5. Initialize an empty co-occurrence sense dictionary, CoocDict.

6. For each target word $t \in T$ do:

6.1 Extract subset $D_t \subset D$ where Target_word = t .

6.2 Extract unique senses (Equation 4):

$$Y_t = \{y_i | (s_i, t_i, y_i) \in D_t\} \quad (4)$$

6.3 Create sense-to-index mapping CoocDict[t] (Equation 5):

$$Y_t \rightarrow \{0, 1, 2, \dots, |Y_t| - 1\} \quad (5)$$

7. Initialize global Label Encoders:

7.1 Create LE_t (Equation 6):

$$T \rightarrow \{0, 1, \dots, |T| - 1\} \quad (6)$$

7.2 Create LE_s (Equation 7):

$$\bigcup_{t \in T} (Y_t \rightarrow \{0, 1, \dots, c - 1\}) \quad (7)$$

8. Encode dataset: For each $(s_i, t_i, y_i) \in D$:

$$8.1 \ t_i^{enc} = LE_t(t_i) \quad (8)$$

$$8.2 \ y_i^{enc} = LE_s(y_i) \quad (9)$$

9. Load pretrained multilingual BERT model and tokenizer.

10. If a cached feature matrix exists, load it; otherwise, compute as follows:

11. Initialize an empty feature matrix X.

12. For each sample $(s_i, t_i, y_i) \in D$ do:

12.1 Tokenize sentence s_i using the mBERT tokenizer.

12.2 Pass tokens through mBERT to obtain hidden states: $h_j^{(l)} \in R^d$

12.3 Compute CLS embedding (Equation 10):

$$e_i = \frac{1}{4} \sum_{r=0}^3 h_{CLS}^{(L-r)} \in R^d \quad (10)$$

12.4 Identify token indices I_i corresponding to target word t_i .

$$I_i = \{j \mid w_j \text{ corresponds to } t_i\}$$

12.5 Compute target embedding (Equation 11):

$$t_i^{bert} = \frac{1}{|I_i|} \sum_{j \in I_i} h_j^{(L)} \in R^d \quad (11)$$

12.6 Concatenate contextual representation (Equation 12):

$$x_i^{bert} = [e; t_i^{bert}] \in R^{2d} \quad (12)$$

12.7 Construct co-occurrence vector:

$$x_i^{cooc} \in R^{mmax}$$

12.8 Extract lemma identity feature (Equation 13):

$$x_i^{lemma} = t_i^{enc} \quad (13)$$

12.9 Construct final feature vector (Equation 14):

$$X_i = [x_i^{bert}; x_i^{cooc}; x_i^{lemma}] \quad (14)$$

12.10 Append X_i to the feature matrix X .

13. Save computed feature matrix X to disk for caching.

14. Construct composite labels for stratified splitting.

For each sample (Equation 15):

$$z_i = (t_i^{enc}, y_i^{enc}) \quad (15)$$

15. Perform Stratified Shuffle Split (80% training, 20% testing).

16. If trained global RF exists, load it; else:

16.1 Train Random Forest Classifier on $RF(X_{train}, y_{train})$

16.2 Save trained model RF, and encoders LE_t , LE_s .

17. Predict senses for test set (Equation 16):

$$\hat{y} = RF(X_{test}) \quad (16)$$

18. Decode predictions using (Equation 17):

$$\hat{y}^{label} = LE_s^{-1}(\hat{y}) \quad (17)$$

19. Compute evaluation metrics: Accuracy, Macro-Precision, Macro-Recall, Macro-F1.

End

3.3. Illustrative Example of the execution Flow

The following is an example that will serve to describe the working of the algorithm that has been designed. This

example highlights the complete process flow of the system, along with explaining how the algorithm decides the sense of the ambiguous word in the sentence.

1. Load the dataset D from disk.

Example:

Sentence	Target_word	Sense
s_1 : આ વર્ગ ના લોકોને વધારે સહાયની જરૂર છે.	t_1 : વર્ગ	y_1 : સમુદ્ધ
s_2 : પાંચ નો વર્ગ શું થાય?	t_1 : વર્ગ	y_2 : ગણિત

2. Normalize all sentences and target words using Unicode normalization and whitespace cleaning.

Actual Sentence: S_1 : આ વર્ગ ના લોકોને વધારે સહાયની જરૂર છે.

Normalized Sentence: આ વર્ગ ના લોકોને વધારે સહાયની જરૂર છે.

Actual Sentence: પાંચ નો વર્ગ શું થાય?

Normalized Sentence: પાંચ નો વર્ગ શું થાય?

3. Extract the set of all sentences S from D .

$S = \{s_1, s_2\}$

4. Extract the set T of all unique target words from D .

$T = \{\text{“વર્ગ”}\}$

5. Initialize an empty co-occurrence sense dictionary
CoocDict={}
6. For $t = \text{"વર્ગ"}$
Extract Subset
 $D_{\text{cooc}} \subset D$
Extract Senses
 $Y_{\text{cooc}} = \{\text{"સમુદ્ધ", "વર્ગભિદ્ધ"}\}$
Create Mapping
 $\text{CoocDict}[\text{"વર્ગ"}] = \{\text{"સમુદ્ધ":0, "વર્ગભિદ્ધ":1}\}$
7. Target Word Encoding (LE_t)
 $LE_t(\text{"વર્ગ"}) = 0$
Global Sense Encoding (LE_s)
 $LE_s(\text{"સમુદ્ધ"}) = 0$
 $LE_s(\text{"વર્ગભિદ્ધ"}) = 1$
8. Encode Dataset

Sentence	t_enc	y_enc
s_1	0	0
s_2	0	1
9. Load Multilingual BERT
Hidden size $d=768$
Number of layers $L=12$
10. If a cached feature matrix exists, load it; otherwise, compute as follows:
11. Initialize an empty feature matrix X .
12. Process Each Sample
 - 12.1 Tokenization
[CLS], આ, વર્ગ, ના, લોકોને, વધારે, સહાયની, જરૂર, છે, ., [SEP]
 - 12.2 Pass tokens through mBERT to obtain hidden representations.
 $h_j^l \in R^{768}$
 - 12.3 Compute CLS Embedding
$$e = \frac{1}{4} \sum_{r=0}^3 h_{CLS}^{(L-r)} \in R^d$$
 - 12.4 Identify token indices I_i corresponding to target word t_i
 - 12.5 Compute target embedding
Example: “આ વર્ગના લોકોને વધારે સહાયની જરૂર છે”
Token index for “વર્ગ” = position 2
 $I_1 = \{2\}$
 $t_1^{bert} = h_2^{(L)}$
 $t_1^{bert} \in R^{768}$
 - 12.6 Contextual representation
 $X_i^{bert} = [e; t] \in R^{2d}$
 $X_1^{bert} = [e; t_1^{bert}]$
 - 12.7 co-occurrence vector
Using the dictionary: $\{\text{"સમુદ્ધ":0, "વર્ગભિદ્ધ":1}\}$
Sentence 1
Sense = “સમુદ્ધ”
 $x_{\text{cooc}} = [1, 0]$
Sentence 2

Sense = “વર્ગભિદ્ધ”

 $x_{\text{cooc}} = [0, 1]$

- 12.8 Extract lemma identity feature

For વર્ગ, $x_1^{lemma} = 0$

- 12.9 Construct final feature vector,
- X_i

 $X_i = [X_i^{bert}; x_i^{cooc}; x_i^{lemma}]$ $X_1 = [1536 \text{ dims}, 1, 0, 0]$ Total dimension: $1536+2+1=1539$

Sentence 1:

 $x_1 = [$ 1536 dims $\rightarrow$ BERT (social context)2 dims $\rightarrow [1, 0]$ (cooc)1 dim $\rightarrow$ lemma ID = 0]

Sentence 2:

 $x_2 = [$ 1536 dims $\rightarrow$ BERT (math context)2 dims $\rightarrow [0, 1]$ (cooc)1 dim $\rightarrow$ lemma ID = 0]

- 12.10 Append
- X_i
- to the feature matrix
- X
- .

 $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ Shape = 2×1539

13. Save computed feature matrix
- X
- to disk for caching.

14. Construct composite labels for stratified splitting.

For each sample $z_i = (t_i^{enc}, y_i^{enc})$ $z_1 = (0, 0)$ $z_2 = (0, 1)$

15. Perform Stratified Shuffle Split (80% training, 20% testing).

Stratification ensures both classes are represented if possible.

16. Train Random Forest Classifier on RF (
- $X_{\text{train}}, y_{\text{train}}$
-)

Model learns mapping: RF: $R^{1539} \rightarrow \{0, 1\}$

Decision boundaries formed using:

Social context vectors $\rightarrow$ Class 0Mathematical context vectors $\rightarrow$ Class 1

17. Predict senses for test set:
- $\hat{y} = RF(X_{\text{test}})$

Suppose test sample = s_2 Model computes: $\hat{y} = RF(X_2)$ Output: $\hat{y} = 1$

18. Decode predictions using
- $\hat{y}^{label} = LE_s^{-1}(\hat{y})$

Decode: $LE_s^{-1}(1) = \text{"વર્ગભિદ્ધ"}$

19. Compute evaluation metrics: Accuracy, Macro-Precision, Macro-Recall, Macro-F1.

4. Results and Discussions

The present section discusses the intelligent analysis of the Gujarati WSD through the application of the context-sensitive hybrid technique. In order to verify the efficiency of the new approach to multi-dimensional feature integration, involving multilingual BERT embeddings, weighted co-occurrence matrices, and lexeme clusters, a lot of experiments have been conducted on the basis of the newly compiled Gujarati Sense dataset.

The classification skills and capacity for generalizing of the global Random Forest ensemble are thoroughly analyzed in terms of standard metrics such as average accuracy, macro-precision, macro-recall, and macro-F1 score calculated under 5-fold cross-validation. The study provides a more in-depth examination of the advantages of the hybrid approach compared to distinct methods. Conclusions drawn about the performance of the method in terms of disambiguation and accuracy in handling regional phrases.

4.1. Experiment Setup and Dataset

In order to verify the performance of the proposed context-aware hybrid approach, an appropriate set of Gujarati corpora was meticulously prepared along with the validation process. In our experiment, we constructed a new Gujarati corpus annotated for testing the context-aware hybrid approach.

This reflects the importance of developing an appropriate framework for WSD. It is made up of 4,400 unique sentences generated from a selected group of 150 ambiguous Gujarati words, which have 426 unique meanings. All experiments were performed using a conventional computing configuration.

The feature extraction uses multilingual BERT. However, the required computing resources for the whole process are quite reasonable since only a standard multi-core CPU, a standard GPU for transformers, and a sufficient amount of

RAM are needed for data processing, model training, and testing activities in general.

The current experiment involved the construction of a new Gujarati corpus, annotated for testing the context-aware hybrid approach. With this fold configuration, the dataset was divided into five equal-sized parts. In each iteration, one part of the dataset is used to train the model, while the other four parts are used to test it.

Basically, then, 5-fold used 80% of the data for training each time and 20% for testing. While training the global RF model, sentence features, in the form of a combination of contextual, statistical, and word-lexical vectors, are computed and used to ingest patterns from manually annotated labels for each word sense.

In addition, a part of the section was left separate and used to test the trained models. Using this test data, which has never been seen before, to evaluate the hybrid model yields an unbiased estimate of how accurately the algorithm disambiguates words in Gujarati sentences that have never been seen before.

In a nutshell, this gives us an idea of how well the model generalizes in a harsh linguistic environment. The details of the reference dataset are shown in Table. Therefore, the next section briefly discusses the results of testing this hybrid classifier on this dataset.

Table 4. Example of Gujarati dataset for WSD

Target Word	Sense	Sentence
કુંભ (Kumbha)	રાશિ (Rāśi/Zodiac)	જ્યોતિષશાસ્ત્રમાં કુંભ રાશિને અગત્યનું સ્થાન છે. (Jyōtiṣaśāstramāṁ kumbha rāśinē agatyanuṁ sthāna chē./ The Aquarius zodiac sign holds an important place in astrology.)
	વાસણ (Vāṣaṇa/Vessel)	કુંભને આંગણે પૂજામાં સ્થાપિત કરવામાં આવ્યો. (Kumbhanē āṁgaṇē pūjāmāṁ sthāpita karavāmāṁ āvyō. / The sacred pot was placed at the entrance for the worship ceremony.)
કાંટો (Kāṁṭō)	અડચણ (Aḍaṇa/Obstacle)	સફળતાની રાહમાં રમેશ મહેશ માટે કાંટો બન્યો. (Saphaḷatānī rāhamāṁ ramēśa mahēśa mātē kāṁṭō banyō. / Ramesh became an obstacle in Mahesh's path to success)
	શૂળ (śūḷa/Thorn)	ગુલાબની ડાળ પર કાંટો છે. (Gulābanī dāḷa para kāṁṭō chē./There is a thorn on the rose stem.
	માછલાં પકડવાનો આંકડો (Māchalām pakadavānō āṁkaḍō/Fishhook)	કાંટો પાણીમાં નાખતાં જ માછલી ફસાઈ ગઈ. (Kāṁṭō pāṇīmāṁ nākhatām ja māchalī phasāī gaī./ The fish got caught as soon as the hook was cast into the water.)
	ઘડિયાળ માંહેની વખત દેખાડતી સળી (Ghaḍiyāḷa māmhēnī vakhata dēkhāḍatī saḷī/The hand that shows the time on a clock.)	ઘડિયાળનો કાંટો આગળ વધતો રહ્યો. (Ghaḍiyāḷanō kāṁṭō āgaḷa vadhatō rahyō./The hands of the clock kept moving forward.)

	ત્રાજવાનો સળિયો (Trājavanō saḷiyō/Balance beam of weighing machine)	કાંટો નીચે ઝૂકતાં સ્પષ્ટ થયું કે વજન વધારે છે. (Kāntō nīcē jhūkatām spaṣṭa thayuṁ kē vajana vadhārē chē./As the pointer dipped, it became clear that the weight was more.)
	તોલવાનું યંત્ર (Tōlavānuṁ yaṁtra/Weighing machine)	કઠોળ અને અનાજને જોખવા માટે કાંટો ખૂબ ઉપયોગી છે. (Kaṭhōḷa anē anājanē jōkhavā māṭē kāntō khūba upayōgī chē./A Weighing machine is very useful for weighing pulses and food.)
કોરું (Kōruṁ)	સૂકું (sūkuṁ/Desiccated)	કોરું લાકડું સળગાવવામાં આવ્યું. (Kōruṁ lākaḍuṁ saḷagāvavāmām āvyuṁ./ Desiccated wood was set on fire.)
	વાપર્યા વગરનું કાપડ (Vāparyā vagaranuṁ kāpaḍa/Unused)	તેણે કોરું કપડું ભેટ આપ્યું. (tēṇē kōruṁ kapaḍuṁ bhēṭa āpyuṁ./He gifted a new, unused cloth.)
	લખ્યા વિનાનું કાગળ (lakhyā vinānuṁ kāgaḷa/ blank sheet of paper)	મેજ પર કોરું કાગળ પડેલું હતું. (mēja para kōruṁ kāgaḷa paḍēluṁ hatuṁ./A blank sheet of paper was lying on the table.)
	ભીનું નહીં તેવું (bhīnuṁ nahīṁ tēvuṁ/Dry)	બેસવાના સ્થાનને કોરું રાખવું જરૂરી છે. (Bēsavanā sthānanē kōruṁ rākhavuṁ jarūrī chē./ It is necessary to keep the seating area dry.)
કડી (kadi)	જોડાણ (jōḍāṇa/ Link)	એક કડી વગર સાંકળ અધૂરી રહે છે. (ēka kaḍī vagara sāṁkaḷa adhūrī rahē chē./A chain remains incomplete without a single link.)
	શહેરનું નામ (śahēranuṁ nāma/ Name of city)	કડી ગુજરાતનું જાણીતું શહેર છે. (Kaḍī gujarātanuṁ jāñītuṁ śahēra chē./Kadi is a well-known city in Gujarat.)
	કવિતાની પંક્તિ (Kavitānī paṁkti/ Stanza of poem)	કવિએ આ કવિતાની દરેક કડીમાં આનંદની અનુભૂતિ વ્યક્ત કરી છે. (Kaviē ā kavitānī darēka kaḍīmām ānaṁdanī anubhūti vyakta karī chē./ The poet has expressed a sense of joy in every stanza of this poem.)
	એક ઘરેણું (ēka gharēṇuṁ/A piece of jewelry)	તેના કાનમાં સોનાની કડી ચમકતી હતી. (Tēnā kānamām sōnānī kaḍī camakatī hatī./ A gold earring was gleaming in her ear.)
	બારણાંની આડી સાંકળ (Bāraṇāṁnī āḍī sāṁkaḷa/ Door latch)	રાત્રે બારણું કડી લગાવીને બંધ કરવું. (Rātrē bāraṇuṁ kaḍī lagāvinē baṁdha karavuṁ./ Lock the door with a latch at night.)

4.2. Performance Analysis of the Proposed Gujarati Word Sense Disambiguation using Context-Aware Hybrid Framework

This subsection evaluates the proposed Gujarati WSD using a context-aware hybrid framework, considering the manually annotated Guj-Sense corpus. The model integrates multilingual BERT embeddings, statistical co-occurrence

features, and lexical identity information and is evaluated via 5-fold cross-validation using a global RF ensemble.

The results, shown in Table, analyze classification performance and stability in terms of accuracy, macro-precision, and robustness in resolving lexical ambiguity in a morphologically rich, low-resource setting.

Table 5. Gujarati word sense disambiguation using a Context-Aware Hybrid framework

Cross-Validation Fold	Measured Value (In %)			
	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Fold-1	97.16	97.68	97.04	96.85
Fold-2	98.52	99.10	98.61	98.58
Fold-3	97.39	98.42	97.78	97.58

Fold-4	97.61	97.71	97.33	97.10
Fold-5	97.04	97.78	96.98	96.91
Mean	97.54	98.14	97.55	97.40
Standard Deviation	97.54 ± 0.53	98.14 ± 0.55	97.55 ± 0.60	97.40 ± 0.64

The results obtained from experiments prove that the proposed approach of feature fusion via concatenation of transformer-based embeddings and statistical as well as lexical vectors provides exceptionally efficient disambiguation.

According to the results obtained during the five-fold cross-validation procedure, the average values of accuracy, macro-precision, macro-recall and macro-F1-score of the global RF algorithm are 97.54%, 98.14%, 97.55% and 97.40%, respectively. An essential aspect of using machine learning approaches in highly morphological languages is to ensure robustness to distributional change.

Fig presents the performance of accuracy during the cross-validation procedure. The accuracy rate demonstrated the best result on Fold-2 (98.52%) and the worst on Fold-5 (97.04%). The narrow area of variance presented with the shaded region denotes that the standard deviation in the measure is only about ± 0.53 , which can be considered as extremely low and makes the classifier highly robust to specific distribution changes, as there is virtually no over-fitting.

The remaining metrics have displayed quite a similar pattern of behavior. Macro-precision's cross-validation progress is shown in Fig, with a mean of 98.14% and standard deviation of ± 0.55 . This figure indicates the extremely low rate of false positives in terms of macro-precision for the automatic disambiguation. Fig presents macro-recall cross-validation performance for the model, where each fold represents one of the five data segments in the training set.

This figure illustrates the performance of the model in associating true senses in the dataset. Two relatively high macro-recall scores of 98.61% and 97.78%, in Fold-2 and Fold-3, respectively, prove that, in most cases, this model can identify word(s) conveys sense.

The contribution of standard deviation to the shaded region is ± 0.60 . Moreover, Fig graphically illustrates macro-F1 score variance intervals. Macro-F1 scored its highest point at 98.58% in Fold-2 and had a relatively small standard deviation of ± 0.64 .

The superior performance of the proposed Context-Aware Hybrid Framework can be attributed to the complementary information captured by its different feature representations. Multilingual BERT generates contextual semantic representations that provide the meaning of a given ambiguous Gujarati word based on its surrounding text. This approach eliminates the shortcomings of traditional methods based solely on surface forms or manually created lexical features.

Furthermore, the statistical co-occurrence representation delivers the extra distributional data about the relation between words of interest and those in context. In addition, the lemma representation keeps the original form and sense of a word. The framework combines such different representations to provide an RF classifier with both contextual and lexical information for differentiating between the meanings of the target words.

This design has great significance in Gujarati WSD since morphological variability, contextual ambiguity, and limited data availability complicate reliable sense discrimination. The enhancement in performance can be attributed to the use of composite representations instead of any of the components in the framework. The integration of complementary representations rather than any single component of the framework.

The empirical evidence supports the main architectural theory employed in this work, that the contextually obtained representation of multilingual BERT improves WSD in Gujarati by strategically merging with statistical word co-occurrence vectors. The overall Macro-F1 score of 97.40% convincingly confirms the efficiency of the ensemble approach in overcoming the problem of inter-class imbalances, successfully capturing less frequent and refined meaning instances of polysemous words as well as their primary meanings.

In all experiments, the second fold of splitting was distinguished by the best results across all measures taken. However, a similar level of effectiveness of the model in all folds shows the limited dependency of the model on its separate parts. To sum up, the cross-validation outcomes approve of the reliability of the method in WSD of Gujarati.

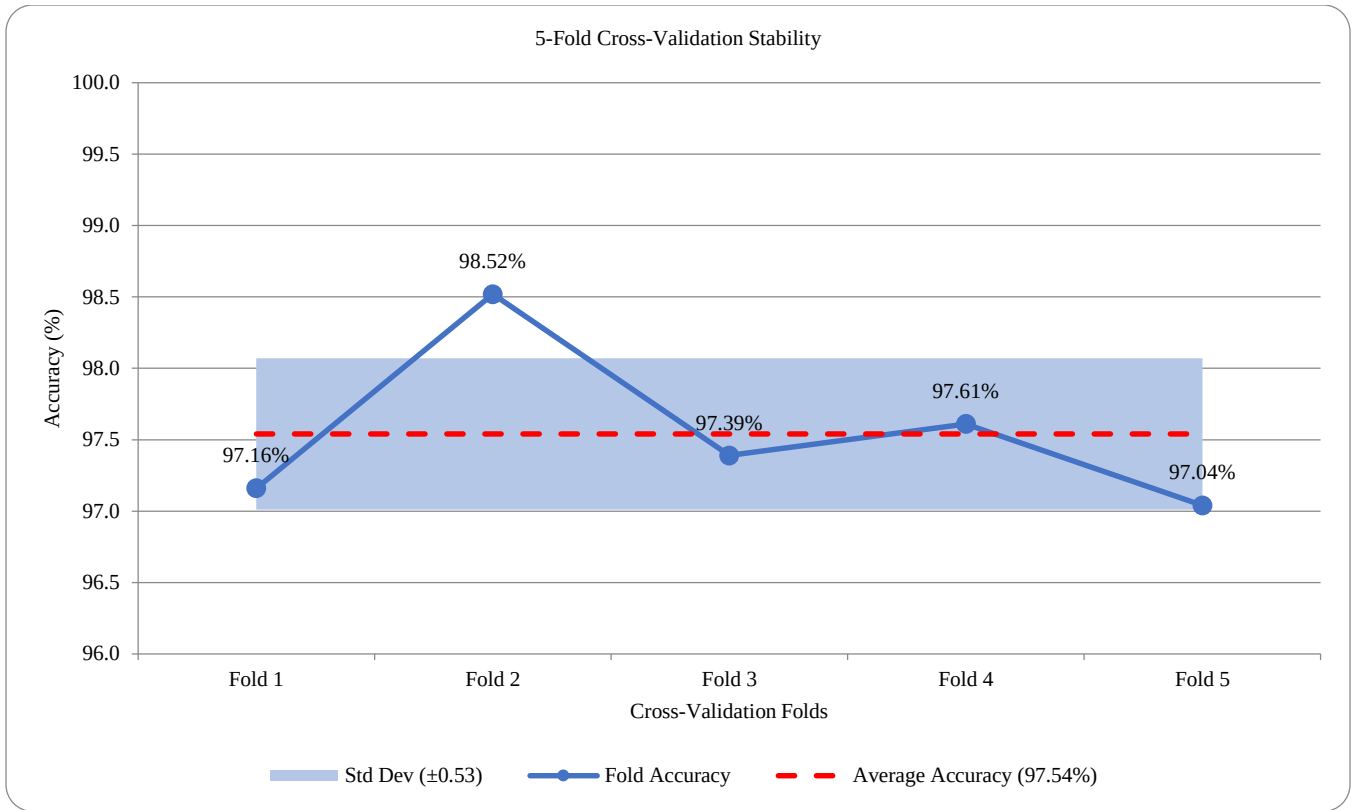


Fig. 3 Cross-validation stability curve illustrating the Accuracy of the proposed Gujarati WSD using a hybrid framework

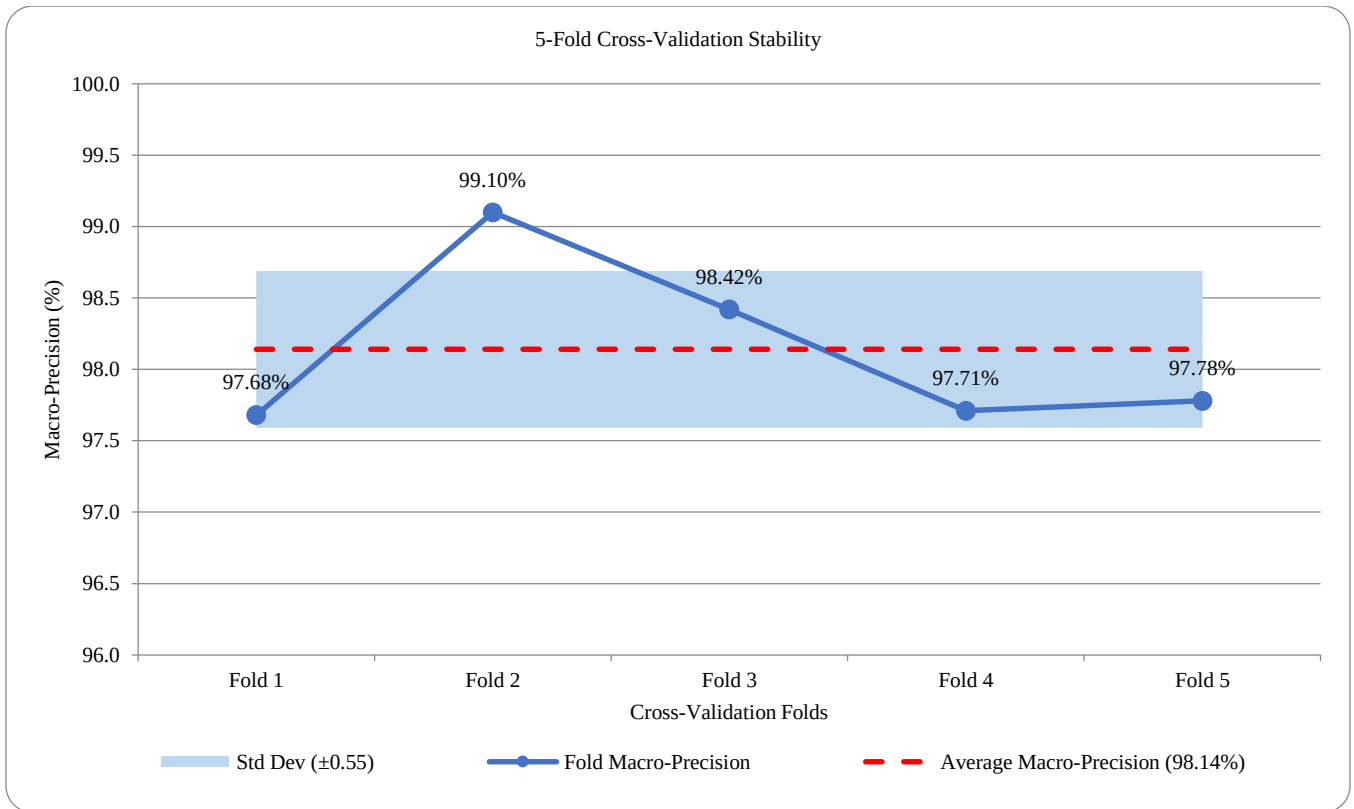


Fig. 4 Cross-validation stability curve illustrating the Macro-Precision of the proposed Gujarati WSD using a hybrid framework

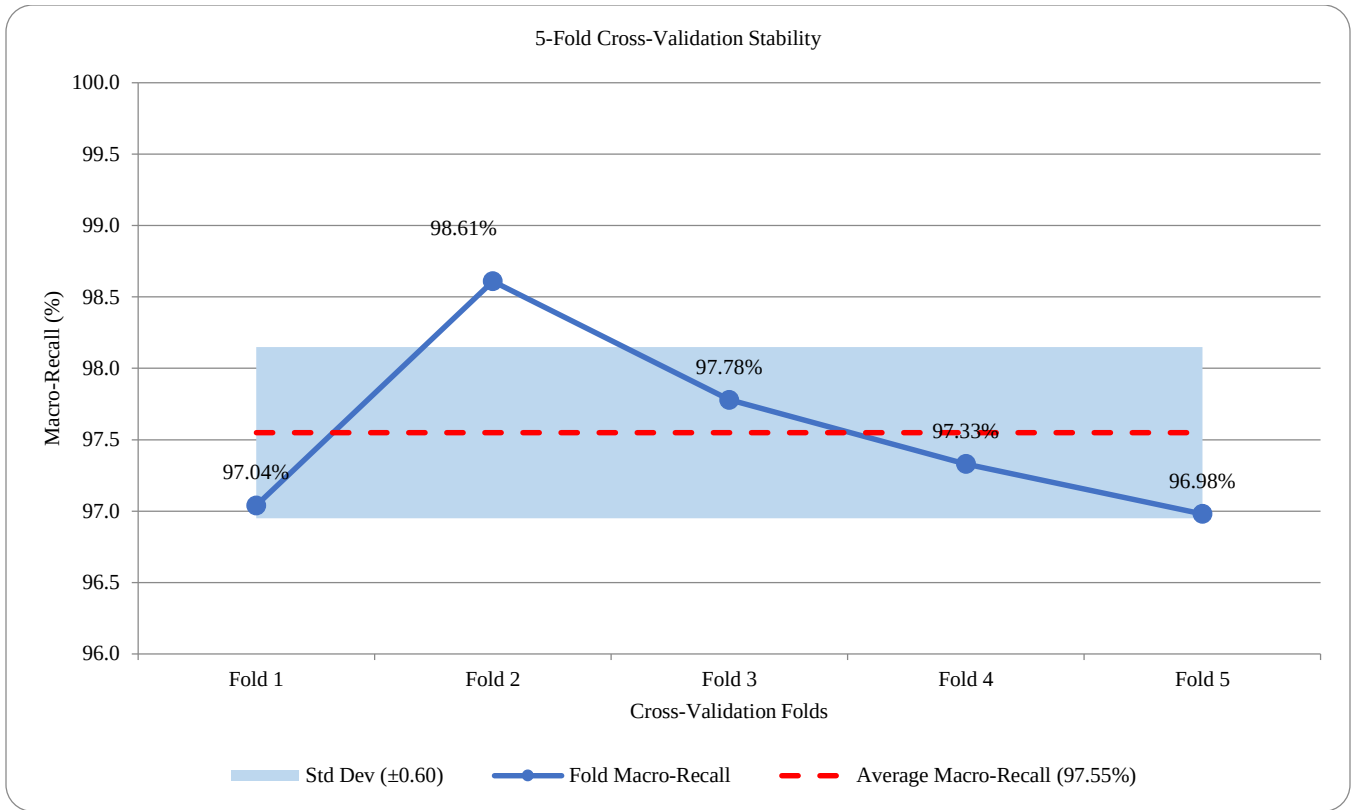


Fig. 5 Cross-validation stability curve illustrating the Macro-Recall of the proposed Gujarati WSD using a hybrid framework

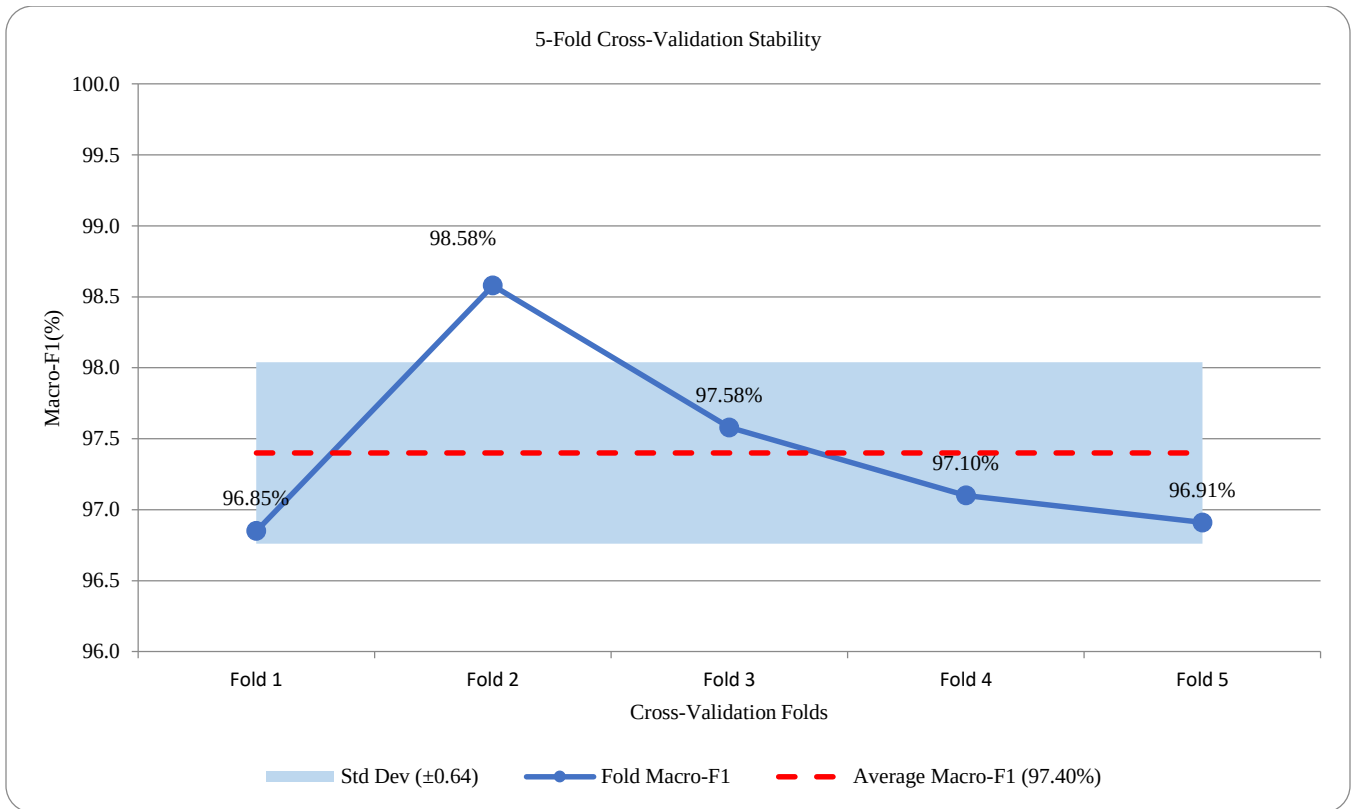


Fig. 6 Cross-validation stability curve illustrating the Macro-F1 of the proposed Gujarati WSD using a hybrid framework

5. Conclusion

This research introduces an extremely robust, context-aware hybrid framework that resolves lexical ambiguity in deeply ambiguous Gujarati with reasonable success. This configuration is a concatenation of deep context embeddings of the multilingual BERT representations and statistical co-occurrence vectors of the statistical matrix and lexical identity mapping in an effort to efficiently represent both global semantics and local syntactic structure. In order to tackle the problem of data sparsity within the context of regional NLP, this research paper employs Guj-Sense, which is a new dataset specifically designed and scaled up for the training and validation of models. This carefully annotated dataset contains manual tagging of 150 distinct polysemous words, which are mapped to 426 different senses based on their appearances in 4,400 sentences contained in the dataset. This model has been rigorously tested using the 5-fold cross-validation approach, and the performance of the model was very impressive. The performance measure attained through this exercise yielded a 97.54% accuracy score along with a Macro-F1 score of 97.40%. The convergence of performance measures implies

that the proposed array-based ensemble Global RF model is reliable against unseen data. The current approach has shown significant discriminative ability and has promising future prospects. To assess the effectiveness of the current highly accurate WSD algorithm for many practical applications and its significance for end-to-end semantic analysis, integration into several other applications, like machine translation systems and cross-lingual information retrieval, will be considered. Besides, it is essential to determine if the application of detailed linguistic features, including syntactic dependencies and POS tagging, may lead to improved accuracy levels. The proposed contextualized hybrid model exhibits the strength of attaining high macro-precision with the help of sentence-level transformer embeddings and co-occurrence vectors. Still, while constraining the semantic window, it confines itself to the analysis within a sentence only. Still, future research can increase the standard by considering WSD at the level of the document or paragraph.

Conflicts of Interest

The authors declare no competing interests.

References

- [1] Eluri Suneetha, Miriyala Kanthirekha, and Dhana Lakshmi Gorle, "Word Sense Disambiguation of Regional Language using Deep Learning," *I-Manager's Journal on Computer Science*, vol. 13, no. 1, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Gerard Escudero Bakx, "Machine Learning Techniques for Word Sense Disambiguation," Polytechnic University of Catalonia, 2006. [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Ajith Abraham et al., "Improvement of Translation Accuracy for the Word Sense Disambiguation System using Novel Classifier Approach," *International Arab Journal of Information Technology*, vol. 21, no. 6, pp. 1128-1142, 2024. [[CrossRef](#)] [[Google Scholar](#)]
- [4] C.P. Chandrika, and Jagadish S. Kallimani, "Word Sense Disambiguation for Indian Regional Language using BERT Model," *Proceedings of Fifth International Conference on Smart Computing and Informatics*, Singapore: Springer Nature Singapore, vol. 2, pp. 127-137, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Sandip S. Patil et al., "Bert and Indowordnet Collaborative Embedding for Enhanced Marathi Word Sense Disambiguation," *ICTACT Journal on Soft Computing*, vol. 13, no. 2, pp. 2842-2849, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Anusha, K. Manjula Shenoy, and Smitha N. Pai, "Embedding Digital News Titles using Wordnet Knowledge and WSD Over Bert Model," *2025 IEEE International Conference on Distributed Computing*, Mangalore, India, pp. 1-6, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Roberto Navigli, "Word Sense Disambiguation: A Survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 1-69, 2009. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Tinghua Wang, Junyang Rao, and Qi Hu, "Supervised Word Sense Disambiguation using Semantic Diffusion Kernel," *Engineering Applications of Artificial Intelligence*, vol. 27, pp. 167-174, 2014. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Abdulgabbbar Saif et al., "Building Sense Tagged Corpus using Wikipedia for Supervised Word Sense Disambiguation," *Procedia Computer Science*, vol. 123, pp. 403-412, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Sunita Rawat, "Supervised Word Sense Disambiguation using Decision Tree," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 4043-4047, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Kaveh Taghipour, and Hwee Tou Ng, "Semi-Supervised Word Sense Disambiguation using Word Embeddings in General and Specific Domains," *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, pp. 314-323, 2015. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Vinto Gujjar et al., "A Literature Survey on Word Sense Disambiguation for the Hindi Language," *Information*, vol. 14, no. 9, pp. 1-25, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Alok Ranjan Pal et al., "Hybrid Approach to Word Sense Disambiguation Combining Supervised and Unsupervised Learning," *International Journal of Artificial Intelligence and Applications*, vol. 4, no. 4, pp. 89-101, 2013. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [14] Tabor Wegi Geleta, and Jara Muda Haro, "Semisupervised Learning-based Word-Sense Disambiguation using Word Embedding for Afaan Oromoo Language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, pp. 1-11, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Tarjni Vyas, and Amit Ganatra, "Gujarati Language Model: Word Sense Disambiguation using Supervised Technique," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2S11, pp. 3740-3744, 2019. [[CrossRef](#)] [[Google Scholar](#)]
- [16] Bhagyada Desai, Krinal Naik, and Preeti Bhatt, "Word Sense Disambiguation in Gujarati Language," *International Journal of Innovative Research in Computer Science and Technology*, vol. 3, no. 1, pp. 44-47, 2015. [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Parth J. Vasoya, and Tarjni Vyas, "A Survey on Word Sense Disambiguation Approaches," *International journal of Trend in Research and Development*, vol. 1, no. 1, pp. 1-3, 2014. [[Publisher Link](#)]
- [18] Zankhana B. Vaishnav, "Gujarati Word Sense Disambiguation using Genetic Algorithm," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 6, pp. 635-639, 2017. [[Google Scholar](#)]
- [19] Tarjni Vyas, and Amit Ganatra, "Gujarati Language: Research Issues, Resources and Proposed Method on Word Sense Disambiguation," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2S11, pp. 3745-3749, 2019. [[CrossRef](#)] [[Google Scholar](#)]
- [20] Manish Sinha et al., "Hindi Word Sense Disambiguation," *Machine Translation*, pp. 1-7, 2000. [[Google Scholar](#)]
- [21] Pranjal Protim Borah, Gitimoni Talukdar, and Arup Baruah, "Assamese Word Sense Disambiguation using Supervised Learning," *2014 International Conference on Contemporary Computing and Informatics*, Mysore, India, pp. 946-950, 2014. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Jumi Sarmah, and Shikhar Kr. Sarma, "Decision Tree based Supervised Word Sense Disambiguation for Assamese," *International Journal of Computer Applications*, vol. 141, no. 1, pp. 42-48, 2016. [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Sailendra Kumar, and Rakesh Kumar, "Word Sense Disambiguation in the Hindi Language: Neural Network Approach," *International Journal of Technical Research and Science*, pp. 72-76, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Nazreena Rahman, and Bhogeswar Borah, "An Unsupervised Method for Word Sense Disambiguation," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6643-6651, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Anh-Cuong Le et al., "Semi-Supervised Learning Integrated with Classifier Combination for Word Sense Disambiguation," *Computer Speech and Language*, vol. 22, no. 4, pp. 330-345, 2008. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Chandrakant D. Kokane, and Sachin D. Babar, "Supervised Word Sense Disambiguation with Recurrent Neural Network Model," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 2, pp. 1447-1453, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Huei-Ling Lai et al., "Supervised Word Sense Disambiguation on Polysemy with Neural Network Models: A Case Study of BUN in Taiwan Hakka," *International Journal of Asian Language Processing*, vol. 30, no. 3, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Anup Kumar Barman et al., "Word Sense Disambiguation Applied to Assamese-Hindi Bilingual Statistical Machine Translation," *Engineering, Technology and Applied Science Research*, vol. 14, no. 1, pp. 12581-12586, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Debapratim Das Dawn et al., "Lexeme Connexion Measure of Cohesive Lexical Ambiguity Revealing Factor: A Robust Approach for Word Sense Disambiguation of Bengali Text," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 12939-12983, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Binod Kumar Mishra, and Suresh Jain, "An Innovative Method for Hindi Word Sense Disambiguation," *SN Computer Science*, vol. 4, no. 6, pp. 1-17, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Debapratim Das Dawn et al., "A Dataset for Evaluating Bengali Word Sense Disambiguation Techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 4057-4086, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Binod Kumar Mishra, and Suresh Jain, "Word Sense Disambiguation for Indic Language using Bi-LSTM," *Multimedia Tools and Applications*, vol. 84, no. 16, pp. 16631-16656, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] Chhaya S. Patil, and Vaishali B. Patil, "A Multilingual Exploration of Word Sense Disambiguation using Transformer Models: Dravidian and Devanagari Languages, 1st ed., Recent Advances in Computing Sciences, CRC Press, pp. 158-160, 2025. [[Google Scholar](#)] [[Publisher Link](#)]
- [34] Sunjae Kwon et al., "Vision Meets Definitions: Unsupervised Visual Word Sense Disambiguation Incorporating Gloss Information," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1583-1598, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Robbel Habtamu Yigzaw, Beakal Gizachew Assefa, and Elefelious Getachew Belay, "A Hybrid Contextual Embedding and Hierarchical Attention for Improving the Performance of Word Sense Disambiguation," *IEEE Access*, vol. 13, pp. 21744-21758, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] Hlaudi Daniel Masethe et al., "Hybrid Transformer-based Large Language Models for Word Sense Disambiguation in the Low-Resource Sesotho sa Leboa Language," *Applied Sciences*, vol. 15, no. 7, pp. 1-33, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [37] Tawseef Ahmad Mir, and Aadil Ahmad Lawaye, "Word Sense Disambiguation Corpus for Kashmiri," *Natural Language Processing*, vol. 31, no. 2, pp. 631-654, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [38] Ali Osman Mohammed Salih et al., "Enhancing a Context-based Supervised Machine Learning Model for Accurate Disambiguation of English Polysemous Word Senses," *Scientific Culture*, vol. 11, no. 4, 1945. [[Google Scholar](#)]
- [39] Shailendra Kumar Patel, Rakesh Kumar, and Anuj Kumar Sirohi, "*BERT for Hindi Word Sense Disambiguation*, 1st ed., Intelligent Computing and Communication Techniques, CRC Press, pp. 747-753, 2025. [[Google Scholar](#)] [[Publisher Link](#)]
- [40] Madhuri Karnik et al., "State of the Art Analysis of Word Sense Disambiguation," *Intelligent Computing for Sustainable Development*, Springer, Cham, pp. 55-70, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [41] Aditi Barai et al., "A Comprehensive Analysis of Word Sense Disambiguation in a Regional Language," *2025 International Conference on Inventive Computation Technologies*, Kirtipur, Nepal, pp. 1260-1265, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [42] Ratul Das, Alok Ranjan Pal, and Diganta Saha, "Unsupervised Approach for Word Sense Disambiguation in Bengali," *Computational Technologies and Electronics*, Springer, Cham, pp. 196-206, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [43] Youddha Beer Singh et al., *A Handbook of Computational Linguistics: Artificial Intelligence in Natural Language Processing*, Bentham Science Publishers, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [44] Lianwang Hao, Tao Zhang, and Huaixin Liang, "A Novel Rotational Causal Random Forest Approach for Word Sense Disambiguation of English Modal Verbs," *SSRN*, pp. 1-9, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [45] Aarti Purohit, and Kuldeep Kumar Yogi, "Enhancing Word Sense Disambiguation for Hindi Agriculture Domain: Feature Engineering and Machine Learning Approaches," *Results in Engineering*, vol. 28, pp. 1-12, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [46] Anirudh S Nair et al., "Improving Word Sense Disambiguation by Adopting Refined Algorithms," *2024 Third International Conference on Trends in Electrical, Electronics, and Computer Engineering*, Bangalore, India, pp. 135-140, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [47] Vivek A. Manwar, and A.B. Manwar, "mBERT: A Query Refinement Model for Marathi Word Sense Disambiguation," *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science*, Bhopal, India, pp. 1-6, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [48] Pawan Makhija, and Sanjay Tanwani, "An Empirical Study on Various Word Sense Disambiguation Techniques in the Biomedical Domain," *International Conference on Recent Advancements and Modernisations in Sustainable Intelligent Technologies and Applications*, Indore, India, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]