Marathi Word Sense Disambiguation through unsupervised K-Means Clustering

Rasika Ransing
Datta Meghe College of Engineering, Navi Mumbai, Maharashtra, India | Vidyalankar Institute of Technology, Mumbai, Maharashtra, India
rasikaransing275@gmail.com (corresponding author)

Archana Gulati
School of Business Management, SVKM's NMIMS University, Navi Mumbai, Maharashtra, India
archana.gulati@nmims.edu

Received: 19 December 2024 | Revised: 25 January 2025 | Accepted: 29 January 2025

 DOI: https://doi.org/10.48084/etasr.9975

ABSTRACT

Word Sense Disambiguation (WSD) is the most crucial Natural Language Processing task and refers to the process of determining the most suitable meaning of a word within its contextual usage. The case of the Marathi language is a bit complicated because it is considered a low-resource language, primarily due to the scarcity of annotated datasets. This study employs an unsupervised machine learning technique using k-means clustering for the disambiguation of Marathi words with more than one meanings without relying on manually labeled data. This disambiguation is accomplished with the help of the context these ambiguous words are used. Instead of implementing k-means clustering concurrently for all 12 words including 42 meanings, it is implemented separately for each word. The number of clusters for each word equals the number of meanings assigned to it. For each word, a Silhouette score is calculated to evaluate the quality of the obtained clustering. In the case of nouns, semantic boundaries were better defined, achieving higher Silhouette scores.

Keywords-unsupervised learning; k-means clustering; word sense disambiguation; Marathi language; natural language processing

I. INTRODUCTION

Word Sense Disambiguation (WSD) is one of the most important tasks of Natural Language Processing (NLP) that deals with the determination of the most suitable meaning of a given word in a given context of a sentence [1, 2]. WSD is important for many NLP functions such as information retrieval, machine translation, and semantic analysis [3, 4]. In languages like Marathi, where there are several meanings of the same word in different contexts, this task is more challenging due to the layers of linguistic and cultural aspects. While there has been considerable work on WSD in languages such as English, WSD studies for Marathi are still few, perhaps due to limited annotated corpora and language resources. In this context, while supervised learning methods show significant potential for addressing WSD, they need a large amount of annotated data, which is difficult to obtain for low-resource languages such as Marathi [5, 6]. Marathi is one of the major Indo-Aryan languages with widespread use in the Indian state of Maharashtra. However, Marathi NLP systems have been relatively sparse compared to more widely spoken languages. The lack of focus on Marathi language processing is a challenge and obstacle to the development of more sophisticated NLP applications for users of this language [6].

WSD is particularly critical in Marathi due to its sole characteristics and wide use of polysemous words. Therefore, without WSD, the systems developed to understand the meaning or interpretation of Marathi may have great difficulty resolving the correct meaning that was assigned to certain phrases, which may then lead to the wrong result for sentiment analysis or machine translation. Considering that the spread of digital input in Marathi is increasing, proper use of WSD can ensure the effectiveness of language processing systems, offering better reliability, easier communication, and better user experiences among people communicating in Marathi. Existing strategies for WSD generally fall under two categories known, respectively, as corpus-based approaches and knowledge-based approaches [7-8].

- Corpus-based methods use a dataset to extract features that provide linguistic information about the contextual components of each sentence. Corpus-based approaches can be categorized into the following techniques:
    - Corpus-based supervised techniques that require a corpus with sense annotations.
    - Unsupervised techniques that do not require corpora annotated with senses.

- Semi-supervised methods can be defined as an amalgamation of both supervised and unsupervised methods. Such approaches use a small amount of sense-tagged corpora along with a large amount of unlabeled corpora.

- Knowledge-based approaches exploit lexical resources such as ontologies, machine-readable dictionaries, and thesauri.

Supervised learning relies significantly on large amounts of labeled data and normally poses a problem in low-resource languages such as Marathi, as the compilation of large annotated corpora is hectic and time-consuming. Unsupervised learning offers several advantages over supervised learning for WSD in Marathi, given the scarcity of annotated datasets and linguistic resources available in the language [2]. Unsupervised methods, primarily clustering algorithms, have been proven to be practical for WSD when using unannotated corpora. K-means clustering, in particular, is a simple yet effective algorithm that identifies clusters by grouping similar data elements and is an attractive strategy for finding meaning distinctions based on unlabeled data [9-10].

There is a need to develop high-quality linguistic resources for Marathi to develop applications such as sentiment analysis, machine translation, etc. Additionally, there is a need to establish standardized benchmarks and evaluation frameworks for the Marathi language. This study presents an unsupervised learning approach using k-means clustering for sense disambiguation of Marathi ambiguous words. This disambiguation is carried out with the help of the context in which the respective ambiguous words are used. Instead of implementing k-means clustering concurrently for all 12 words including 42 senses, k-means clustering is implemented separately for each word. The number of clusters is determined based on the multiple senses linked to each word. The Silhouette score is then calculated for each word to assess the quality of the clustering.

II. RELATED WORKS

In [10], a WSD system for Malayalam was presented, a language spoken in Kerala, India, utilizing a corpus gathered from various online documents. For each potential meaning of ambiguous terms, a limited collection of training samples (seed sets) was selected to represent that meaning. Collocations and most co-occurring words were regarded as training instances. The seed set extension module enhanced the seed set by including the most analogous terms to its constituents. These extended sets function as sensory clusters. The sense cluster most analogous to the input text context is considered the meaning of the target word. This approach generates sense clusters using seed sets. Subsequently, the most analogous meaning was chosen for the uncertain term based on the similarity between the provided input text and the sense clusters. This technique achieved an accuracy of 72%.

In [11], an unsupervised method was used to perform WSD in the Bengali language. The initial phase of this investigation involved executing phrase clustering using the maximum entropy method, with human interaction providing annotations for the clusters' intrinsic meanings, since these sense-tagged clusters may function as sense inventories for subsequent experiments. In the next phase, when the test data necessitate disambiguation, the cosine similarity measure is employed to assess the proximity of the test data to the initially sense-tagged clusters. The least distance of the test data from a specific sense-tagged cluster assigns the same sense to the test data as that of the designated cluster. This method was considered the standard approach, achieving a 35% accuracy rate in the WSD challenge. Subsequently, two enhancements were applied to this foundational strategy: (a) Principal Component Analysis (PCA) utilized on the feature vector, achieving 52% accuracy in the WSD test, and (b) context expansion of sentences employing Bengali WordNet along with PCA, resulting in 61% accuracy in the WSD challenge. The datasets employed in this work were derived from the Bengali corpus, established under the Technology Development for Indian Languages (TDIL) project of the Government of India, with the lexical knowledge base, namely the Bengali WordNet [11, 12].

In [13], an unsupervised learning approach was proposed for WSD, utilizing semantic information to correctly assign the intended sense to ambiguous words. This method employed latent semantic analysis, extracting intrinsic semantic relations among words by analyzing the word distribution in large corpora. Latent semantic analysis helps to aggregate word occurrences of similar meaning into a high-dimensional semantic representation. This method captures the semantic links and subtleties well and classifies fuzzy words into conglomerates that represent their multiple meanings. This study discussed SVD in reducing the dimensionality of the semantic space to ensure better computational efficiency as well as discovering relevant semantic features. The LSA-based WSD method turned out to be competitive with the traditional supervised approaches, especially in situations where labeled data are scarce or unavailable.

In [14], an unsupervised technique for WSD in the Bengali language was proposed. This approach consisted of two successive subtasks. The initial task involved categorizing Bengali sentences into specific clusters, each containing sentences of analogous meaning. The subsequent task involved labeling these clusters with their intrinsic meanings, facilitated by a linguistic expert, as these sense-tagged clusters can serve as a knowledge reference for the WSD task. Clustering was performed with the Weka 3.6.13 tool. The test phrases were derived from the Bengali text corpus established by the Technology Development for Indian Languages (TIDL) project of the Government of India. This study presented type-based and token-based distribution methods for grouping Bengali sentences. The type-based technique used a feature vector including cooccurring terms of a target word within a sentence, and the token-based method also took into account the synsets of the collocating words. The synsets of the collocation terms were extracted from Bengali WordNet, created at ISI, Kolkata.

III. DATASET

Marathi is a language with significantly fewer resources. Semantic tools are not yet fully developed for the Marathi language. The authors did not find a suitable dataset to evaluate the performance of the algorithm, as existing datasets for the Marathi language consist of a few sentences with ambiguous

words. Therefore, a new dataset was created to evaluate the performance of identifying ambiguous words using the k-means algorithm. The ambiguous words were identified from Marathi WordNet, which is a lexical database for Marathi developed by CFILT, IIT Bombay [15]. This work focuses on the disambiguation of 12 ambiguous Marathi words selected manually from this WordNet. These 12 ambiguous words have a total of 42 meanings and comprise 2 verbs, 2 adverbs, 2 adjectives, and 6 nouns. Sentences containing these ambiguous words were collected from various sources, such as Marathi news articles, books, literature, blogs, social media posts, Wikipedia pages, etc., using web scrapers. The web scrapers extract sentences containing the selected ambiguous words. All occurrences of the ambiguous words were then identified in the collected corpus, and the entire sentence around the word was retained. In case of unavailability of a sufficient number of sentences per meaning, a few sentences were created manually also. The appropriate sense of ambiguous words was manually labeled in each sentence. A total of 650 sense-annotated Marathi sentences were compiled in this manner. The dataset quality was ensured by validating it using consistency checks to verify that the sentences were labeled correctly. Wherever required, the sense definitions were refined to account for the appropriate senses. The dataset was organized in columns such as "Ambiguous word", "Number of senses", "Sense", "Sense ID", "Sentence ID", and "Sentences" in a CSV file for easy text processing. There are approximately 15-16 sentences per sense of every ambiguous word. The training set consisted of 500 Marathi sentences, and the performance of the k-means algorithm was evaluated on the remaining 150 test sentences. Table I shows a sample set of sentences from the dataset.

TABLE I. SAMPLE SENTENCES FROM THE DATASET

Ambiguous word	Number of meanings	Sense	Sense ID	Sentence ID	Sentences
वर	5	उंच ठिकाणी	1	1	विमान वर उडत होते.
			1	2	पर्वताच्या वर एक प्राचीन मंदीर आहे.
		एखाद्या वस्तूच्या आधारावर	2	3	पुस्तक शेल्फवर ठेवा.
			2	4	घड्याळ भिंतीवर आहे.
		जास्त वा अधिक	3	5	त्याचे गुण 90 च्या वर आले.
			3	6	तापमान 40 डिग्रीच्या वर आहे.
		च्या मदतीने किंवा आधाराने	4	7	तो आपल्या आई वर विश्वास ठेवतो.
			4	8	इंटरनेट वर मी माहिती शोधली.
		अगोदर वा आधी (लिखाणामध्ये)	5	9	वर लिहिलेल्या अटी वाचा.
			5	10	वर सांगितलेल्या वेळेत भेटा.

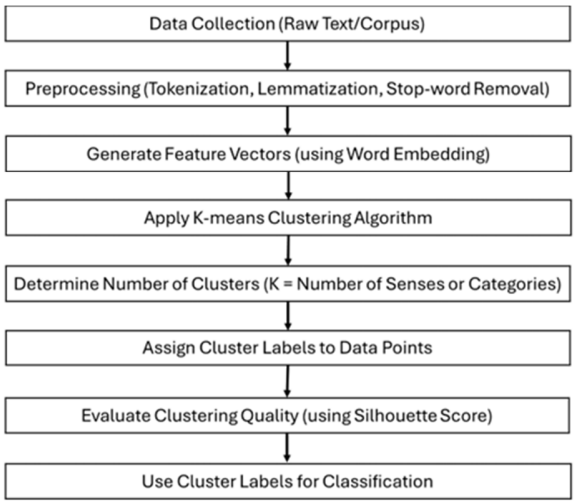

Fig. 1. The proposed method of k-means clustering for Marathi WSD.

IV. PROPOSED METHOD

Figure 1 illustrates the process of using a k-means clustering-based unsupervised learning strategy for WSD in the Marathi language. This includes the following steps:

1. Data collection: This involves collecting a large corpus of Marathi text. This unprocessed text acts as the foundation for recognizing the different environments in which words occur, which is very critical in separating their meanings.
2. Preprocessing: This involves the transformation of raw text into an appropriate format for analysis. A piece of text is broken down into tokens (tokenization), words are reduced to their base forms (lemmatization), and commonly used noninformative words are removed (stop-word elimination). This process ensures that only relevant linguistic features are passed for further analysis.
3. Feature vector generation with word embedding: After preprocessing, every word, or an ambiguous occurrence of a word, is mapped to its corresponding feature vector by word embeddings. The semantic meaning of words represented using word embeddings in a multi-dimensional space aids the analysis of contexts and similarities between words.
4. Apply k-means clustering: These feature vectors are fed into the k-means clustering algorithm, which places instances of ambiguous words in clusters based on similarity in context vectors. Ideally, each cluster should correspond to one distinct sense of the ambiguous word.
5. Determine the number of clusters *k* (number of meanings or categories): The number of clusters k is equal to the number of possible meanings or senses of the ambiguous word(s). In this step, a suitable *k* value is determined, which can be achieved using several methods or by knowing from prior sense definitions.
6. Tagging data points with labels: After clustering, all occurrences of the ambiguous word are tagged by using

the appropriate cluster to which the occurrence belongs. In this way, the labels correspond to the various senses of the word. Consequently, given the context, the algorithm may predict the correct meaning of the word.

7. Evaluation of cluster quality - Silhouette score: The Silhouette score indicates how well the word instances fit into their cluster. The higher the Silhouette score, the better the accuracy of the cluster. A high silhouette score suggests that the clusters represent distinct senses effectively.
8. Use cluster labels for classification: Finally, cluster labels are used to classify the meanings of words in new contexts. This classification enables the system to disambiguate word meanings accurately in future text data, providing a functional WSD system for Marathi.

These steps are applied to the dataset consisting of 650 sentences for 12 ambiguous words. The number of senses (k) for each word is determined based on Marathi WordNet. The sentences in this dataset are preprocessed before applying the k-means algorithm. The following steps are carried out for preprocessing:

i. Tokenization: The sentences are split into words.

ii. Stop-word removal: Frequent and meaningless Marathi stop-words are removed from every sentence. (e.g., "आणि," "तो," "ते").

iii. Lemmatization: The remaining content words are reduced to their base form using morphological analysis.

For each target ambiguous word, a window of n-words around it is selected to capture its context. The extracted contexts are converted into numerical representations. Word2Vec is used to convert the words in the text to vector representations. Each word is represented as a dense vector in a high-dimensional space (d). The context embedding for each target ambiguous word is calculated by averaging the embeddings of all words in its context window. These context vectors are then clustered to disambiguate word senses. k centroids ($\mu_1, \mu_2,\ldots,\mu_k$) are chosen, where k is the number of senses for the target ambiguous word. The distance between each context vector (x_j) and all centroids (μ_i) is calculated using the Euclidean distance [16]:

$$d(x_j, \mu_i) = \sqrt{\sum_{l=1}^{d}(x_{jl} - \mu_{il})^2}$$

Each context vector is assigned to the nearest centroid:

$$c_j = \arg\min_i ||x_j - \mu_i||^2$$

The centroids are recalculated as the mean of all vectors assigned to each cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

The process of distance calculation, cluster assignment, and centroid update are repeated until the centroids stabilize or the change in the objective function (J) falls below a threshold:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x_j - \mu_i||^2$$

While developing the dataset, the sentences with ambiguous words were annotated with appropriate senses based on their context. The meanings predicted by the applied k-means clustering algorithm were then compared with the actual meanings manually annotated in the dataset. The quality of clustering and disambiguation of the meanings is then evaluated using the Silhouette score, which measures how well data points are clustered together using parameters such as cohesion and separation. Cohesion indicates how similar a data point is to other data points in its cluster. Separation indicates how different a data point is from data points in other clusters.

$$Silhoutte\ score = \frac{b-a}{\max(a,b)}$$

where a is the average intra-cluster distance and b is the average nearest-cluster distance. The Silhouette score ranges from -1 to 1 and higher values indicate better clustering.

V. RESULTS AND DISCUSSIONS

The k-means clustering for the word "वर" works as described below. There are 5 possible meanings for this word, i.e., k = 5. The sample sentences are:

- विमान वर उडत होते → The airplane was flying above. (meaning: उंच ठिकाणी → above),
- बाटली टेबलच्या वर आहे → The bottle is on the table. (meaning: एखाद्या वस्तूच्या आधारावर → on),
- त्याचे गुण 90 च्या वर आले → His marks were more than 90. (meaning: जास्त वा अधिक → more than),
- तो आपल्या आई वर विश्वास ठेवतो → He trusts upon his mother (meaning: च्या मदतीने किंवा आधाराने → upon),
- वर नमूद केलेल्या सूचनांचे अनुसरण करा → Follow the aforementioned instructions (meaning: अगोदर वा आधी (लिखाणामध्ये) → aforementioned).

The sentences in the dataset are preprocessed as follows. An instance of preprocessing the sentence "बाटली टेबलच्या वर आहे" (The bottle is on the table) follows:

- Tokenization: "बाटली टेबलच्या वर आहे" → ["बाटली", "टेबलच्या", "वर", "आहे"].
- Stopword removal: Non-informative words like "आहे" are removed. ["बाटली", "टेबलच्या", "वर", "आहे"] → ["बाटली", "टेबलच्या", "वर"].
- Lemmatization: ["बाटली", "टेबलच्या", "वर"] → ["बाटली", "टेबल", "वर"].

The text is converted into numerical vectors. Word embeddings for the words of the given sentence are:

- "बाटली" → [0.2, 0.4, 0.1],
- "टेबल" → [0.5, 0.3, 0.6],
- "वर" → [0.3, 0.5, 0.2].

The average vectors of the words in the sentence is calculated.

Context vector (v):

$$v = \frac{1}{3}([0.2, 0.4, 0.1] + [0.5, 0.3, 0.6] + [0.3, 0.5, 0.2])$$

$$v = [0.33, 0.4, 0.3]$$

This is repeated for all sentences to obtain the context vectors for "वर".

As the number of senses for "वर" is initialized as $k = 5$, five centroids $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ are randomly initialized in the vector space. Each context vector is assigned to the nearest centroid based on Euclidean distance.

For the context vector for "बाटली टेबलच्या वर आहे" ($x_1 = [0.33, 0.4, 0.3]$). Assume the centroids as:

- $\mu_1 = [0.5, 0.5, 0.6]$ (Cluster 1: "above"),
- $\mu_2 = [0.3, 0.35, 0.28]$ (Cluster 2: "on"),
- $\mu_3 = [0.6, 0.7, 0.6]$ (Cluster 3: "more than"),
- $\mu_4 = [0.5, 0.4, 0.32]$ (Cluster 4: "upon"),
- $\mu_5 = [0.5, 0.6, 0.5]$ (Cluster 5: "aforementioned").

The distances are calculated as:

$$d(x_1, \mu_1) = \sqrt{(0.33 - 0.5)^2 + (0.4 - 0.5)^2 + (0.3 - 0.6)^2} = 0.359$$
$$d(x_1, \mu_2) = \sqrt{(0.33 - 0.3)^2 + (0.4 - 0.35)^2 + (0.3 - 0.28)^2} = 0.062$$
$$d(x_1, \mu_3) = \sqrt{(0.33 - 0.6)^2 + (0.4 - 0.7)^2 + (0.3 - 0.6)^2} = 0.503$$
$$d(x_1, \mu_4) = \sqrt{(0.33 - 0.5)^2 + (0.4 - 0.4)^2 + (0.3 - 0.32)^2} = 0.171$$
$$d(x_1, \mu_5) = \sqrt{(0.33 - 0.5)^2 + (0.4 - 0.6)^2 + (0.3 - 0.5)^2} = 0.33$$

These steps are repeated for all vectors for cluster assignment.

The centroids are then recalculated. For cluster 1 ("above"),

Context vectors: [0.33,0.4,0.3], [0.6,0.7,0.8]

New centroid: $\mu_1 = \frac{1}{2}([0.33,0.4,0.3] + [0.6,0.7,0.8]) = [0.465,0.55,0.55]$

This process is repeated until the centroid converges.

The trained unsupervised k-means clustered model is then tested on unseen sentences. The new sentence is preprocessed, and then, a context vector is generated. The distance of this vector to the centroids is calculated and the sentence is assigned to the nearest cluster.

Consider a new sentence "झाडाच्या वर पक्षी आहे" ("There is a bird on the tree"). This sentence is preprocessed:

- Tokenization: "झाडाच्या वर पक्षी आहे" → ["झाडाच्या", "वर", "पक्षी", "आहे"].
- Stopword removal: ["झाडाच्या", "वर", "पक्षी", "आहे"] → ["झाडाच्या", "वर", "पक्षी"].
- Lemmatization: ["झाडाच्या", "वर", "पक्षी"] → ["झाड", "वर", "पक्षी"].

The context vector is generated using Word2Vec and the average vector for the words in the sentence is calculated:

Context: ["झाड", "वर", "पक्षी"]

Vectors are as follows:

- "झाड" → [0.3, 0.5, 0.2],
- "वर" → [0.4, 0.2, 0.5],
- "पक्षी" → [0.4, 0.4, 0.3].

The average context vector (v) is:

$$v = \frac{1}{3}([0.3, 0.5, 0.2] + [0.4, 0.2, 0.5] + [0.4, 0.4, 0.3])$$

$$v = [0.37, 0.37, 0.33]$$

Calculating distances to the centroids:

New sentence vector $v = [0.37, 0.37, 0.33]$

Distances to the vectors are as follows:

$$d(v, \mu_1) = \sqrt{(0.37 - 0.5)^2 + (0.37 - 0.5)^2 + (0.33 - 0.6)^2} = 0.327$$
$$d(v, \mu_2) = \sqrt{(0.37 - 0.3)^2 + (0.37 - 0.35)^2 + (0.33 - 0.28)^2} = 0.088$$
$$d(v, \mu_3) = \sqrt{(0.37 - 0.6)^2 + (0.37 - 0.7)^2 + (0.33 - 0.6)^2} = 0.484$$
$$d(v, \mu_4) = \sqrt{(0.37 - 0.5)^2 + (0.37 - 0.4)^2 + (0.33 - 0.32)^2} = 0.134$$
$$d(v, \mu_5) = \sqrt{(0.37 - 0.5)^2 + (0.37 - 0.6)^2 + (0.33 - 0.5)^2} = 0.314$$

The sentence is assigned to the cluster with the smallest distance, i.e., $d(v, \mu_2) = 0.088$, cluster 2 ("on"). In this way, the k-means clustering disambiguates the ambiguous words concerning the context.

The proposed method cannot be directly compared with other studies due to the absence of standard datasets in Marathi and the limited works focusing on WSD for this language. Unlike widely studied languages, such as English, for which large annotated corpora and benchmark datasets exist, Marathi lacks such resources, making it challenging to evaluate and compare the performance of different WSD methods. This scarcity of standardized datasets hinders consistent benchmarking and impedes the replication of experiments. Additionally, the relatively little research on Marathi WSD means there is no established baseline or widely accepted methodology for this task. As a result, the effectiveness of the proposed unsupervised approach in this study cannot be

measured against other methods in a meaningful way, highlighting the need for resource development and further research in Marathi language processing. In [14], clustering for Bengali WSD obtained an accuracy of 63%. In [17, 18], a Support Vector Machine (SVM) supervised algorithm and unsupervised approach were proposed using word embeddings for Marathi WSD. The accuracy of k-means clustering, SVM, and the unsupervised approach using word embeddings for disambiguating Marathi words was 51%, 53%, and 69% respectively.

TABLE II. EVALUATION OF K-MEANS CLUSTERING FOR AMBIGUOUS WORDS

Sr. No.	Ambiguous word	POS tag	Number of meanings	Silhouette score
1	वर	Adverb	5	0.212
2	बरोबर	Adverb	3	0.281
3	हलका	Adjective	4	0.228
4	मंद	Adjective	4	0.293
5	फोडणे	Verb	4	0.264
6	उडणे	Verb	5	0.295
7	वार	Noun	4	0.295
8	गंध	Noun	2	0.206
9	पाठ	Noun	2	0.286
10	अंक	Noun	3	0.342
11	अर्थ	Noun	3	0.3
12	गुण	Noun	3	0.286

Table II shows the results for the WSD task applied to Marathi words, focusing on the quality of clustering measured by the Silhouette score, which ranged from 0.206 to 0.342. This is a relatively low to moderate range, indicating that, while some clusters are somewhat distinct, there is still overlap or ambiguity in clustering for many words. A higher silhouette score (closer to 1) would indicate more clearly marked clusters with distinct separations between meanings, but this range suggests moderate clustering quality. The words "अंक" (noun) with a score of 0.342 and "अर्थ" (noun) with a score of 0.3 have the highest silhouette scores. This could also reflect that nouns may have clearer semantic boundaries between senses compared to other parts of speech in Marathi. Words such as "मंद" (noun) with a score of 0.206 and "वर" (adverb) with 0.212 have the lowest silhouette scores. This suggests that the clusters for these words are not well-separated, indicating high overlap between senses. Lower scores for adverbs and adjectives, such as "वर" and "हलका" (0.228), imply that these parts of speech may have more subtle or context-dependent sense distinctions that are harder to capture with k-means clustering. As an example, nouns often refer to concrete objects, and thus the segregation between word meanings is quite simply defined. For the case of adverbs and adjectives, whose meanings are more abstract or variable, it can lead to clustering problems, and this is reflected in the Silhouette scores.

VI. CONCLUSION

This study presents an unsupervised k-means clustering approach to word sense disambiguation in Marathi. The results indicate that there is a possibility of applying such clustering-based techniques to WSD tasks for low-resource languages that contain fewer annotated datasets. This study used raw text data and embedding techniques to cluster several meanings of ambiguous words in the Marathi language. This method was effective in disambiguating polysemous words with minimal supervision. Future work will involve more sophisticated unsupervised methods, including graph-based algorithms. Furthermore, this study observed variability across parts of speech in clustering effectiveness, as nouns had high Silhouette scores with less overlap and clearer contextual boundaries, whereas adverbs and adjectives exhibited a kind of ambiguity through their subtle meaning shifts. This is in line with the inherent linguistic complexities of languages such as Marathi, where context plays a vital role in deciding how meaning should be disambiguated. The Silhouette scores being roughly in the middle indicate that there is scope for improvement in terms of clustering quality, especially for words that may not have any different meanings closely related. Advanced clustering algorithms, such as hierarchical or density-based clustering, could be used to increase the distinguishing power between meanings. Advanced contextual embeddings or syntactic dependencies could also be incorporated into a future work stream. Further dataset expansion and description of feature extraction can also lead to improved WSD in Marathi.

REFERENCES

[1] X. Zhang *et al.*, "Word Sense Disambiguation by Refining Target Word Embedding," in *Proceedings of the ACM Web Conference 2023*, New York, NY, USA, Dec. 2023, pp. 1405–1414, https://doi.org/10.1145/3543507.3583191.

[2] P. Jha, S. Agarwal, A. Abbas, and T. J. Siddiqui, "A Novel Unsupervised Graph-Based Algorithm for Hindi Word Sense Disambiguation," *SN Computer Science*, vol. 4, no. 5, Sep. 2023, Art. no. 675, https://doi.org/10.1007/s42979-023-02116-1.

[3] M. Alian and A. Awajan, "Arabic word sense disambiguation using sense inventories," *International Journal of Information Technology*, vol. 15, no. 2, pp. 735–744, Feb. 2023, https://doi.org/10.1007/s41870-022-01147-w.

[4] A. K. Barman, J. Sarmah, S. Basumatary, and A. Nag, "Word Sense Disambiguation applied to Assamese-Hindi Bilingual Statistical Machine Translation," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12581–12586, Feb. 2024, https://doi.org/10.48084/etasr.6342.

[5] C. D. Kokane, S. D. Babar, P. N. Mahalle, and S. P. Patil, "Word Sense Disambiguation: Adaptive Word Embedding with Adaptive-Lexical Resource," in *Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023*, 2023, pp. 421–429, https://doi.org/10.1007/978-981-99-3878-0_36.

[6] P. Lahoti, N. Mittal, and G. Singh, "A Survey on NLP Resources, Tools, and Techniques for Marathi Language Processing," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 2, Sep. 2022, https://doi.org/10.1145/3548457.

[7] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, "Recent Trends in Word Sense Disambiguation: A Survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, Canada, Aug. 2021, pp. 4330–4338, https://doi.org/10.24963/ijcai.2021/593.

[8] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, Oct. 2009, Art. no. 10, https://doi.org/10.1145/1459352.1459355.

[9] D. Ustalov, D. Teslenko, A. Panchenko, M. Chernoskutov, C. Biemann, and S. P. Ponzetto, "An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages." arXiv, Apr. 27, 2018, https://doi.org/10.48550/arXiv.1804.10686.

[10] K. P. S. Sankar, P. C. R. Raj, and V. Jayan, "Unsupervised Approach to Word Sense Disambiguation in Malayalam," *Procedia Technology*, vol. 24, pp. 1507–1513, Jan. 2016, https://doi.org/10.1016/j.protcy.2016.05.106.

[11] A. R. Pal, D. Saha, S. Naskar, and N. S. Dash, "Word sense disambiguation in Bengali: A lemmatized system increases the accuracy of the result," in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, Kolkata, India, Jul. 2015, pp. 342–346, https://doi.org/10.1109/ReTIS.2015.7232902.

[12] A. R. Pal and D. Saha, "Word Sense Disambiguation in Bengali language using unsupervised methodology with modifications," *Sādhanā*, vol. 44, no. 7, Jun. 2019, Art. no. 168, https://doi.org/10.1007/s12046-019-1149-2.

[13] D. I. Martin, M. W. Berry, and J. C. Martin, "Semantic Unsupervised Learning for Word Sense Disambiguation," in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds. Springer International Publishing, 2020, pp. 101–120.

[14] A. R. Pal and D. Saha, "Word sense disambiguation in Bengali: An unsupervised approach," in *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, Feb. 2017, pp. 1–5, https://doi.org/10.1109/ICECCT.2017.8117901.

[15] L. Popale and P. Bhattacharyya, "Creating Marathi WordNet," in *The WordNet in Indian Languages*, N. S. Dash, P. Bhattacharyya, and J. D. Pawar, Eds. Springer, 2017, pp. 147–166.

[16] J. Qi, Y. Yu, L. Wang, and J. Liu, "K*-Means: An Effective and Efficient K-Means Clustering Algorithm," in 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, USA, Oct. 2016, pp. 242–249, https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.46.

[17] R. Ransing and A. Gulati, "Word Sense Disambiguation for Marathi language using Supervised Learning," in *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, Goa University, Goa, India, Sep. 2023, pp. 754–759. [Online]. Available: https://aclanthology.org/2023.icon-1.76/.

[18] R. Ransing and A. Gulati, "Unsupervised Word Sense Disambiguation for Marathi language using Word Embeddings," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 3, pp. 1374–1380, Mar. 2024.