SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 11, No. 4, (2025), pp. 1945-1955
Open Access. Online & Print

DOI: 10.5281/zenodo.11425252

# ENHANCING A CONTEXT-BASED SUPERVISED MACHINE LEARNING MODEL FOR ACCURATE DISAMBIGUATION OF ENGLISH POLYSEMOUS WORD SENSES

**Ali Osman Mohammed Salih[1*], Alwalid Bashier Gism Elseed Ahmed[2], Hatim Mohamed Ahmed Osman[3], Mahmoud Khalifa[4], Ali Siddig Ali Ahmed[5], Nissrein Babiker Mohammed Babiker[6], and Fatima Rayan Awad Ahmed[7]**

*[1]Department of Information Systems and Cyber Security, College of Computing and Information Technology, University of Bisha, University of Bisha, Bisha 61922 P.O. Box: 551, Saudi Arabia 1,6; : aomohammed@ub.edu.sa, nbbabaker@ub.edu.sa*
*[2]Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, P.O. Box: 551, Bisha 67714, Saudi Arabia 2,4,5; Alwlidbasheer@ub.edu.sa, mkhalifa@ub.edu.sa, asiddig@ub.edu.sa,*
*[3]Department of English Language, College of Education, University of West Kordofan, Sudan 3; brother2009@gmail.com*
*[4]Department of Computer Science Department, College of Computer Engineering and Science, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia P.O. Box: 16273, Saudi Arabia 7; fe.ahmed@psau.edu.sa*

*Received: 11/11/2025*
*Accepted: 18/12/2025*

*Corresponding Author: Ali Osman Mohammed Salih (aomohammed@ub.edu.sa)*

## ABSTRACT

***Word Sense Disambiguation (WSD) remains a fundamental challenge in natural language processing, especially when dealing with polysemous words that carry multiple meanings. This study investigates a lightweight, interpretable approach to WSD by relying solely on textual definitions rather than complex contextual embeddings. A carefully constructed dataset containing 256 English polysemous words, each associated with two distinct definitions, was used to frame the task as a binary classification problem. Three traditional supervised machine learning algorithms, Logistic Regression, Support Vector Machine (SVM), and Random Forest, were applied. Definitions were transformed into numerical vectors using Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction. Model performance was evaluated using accuracy, precision, recall, and F1-score, with SVM yielding the highest results across all metrics. The study highlights that conventional machine learning techniques remain effective and efficient in controlled, low-resource scenarios, offering a transparent alternative to deep learning models like BERT. Error analysis revealed that most misclassifications stemmed from semantic proximity between definitions and insufficient contextual depth. While the study acknowledges the limitations of dataset size and lack of broader context, it demonstrates the value of traditional methods in specific applications such as education and digital lexicography. Future work is recommended to expand the dataset and explore hybrid models that balance simplicity, accuracy, and interpretability.***

**KEYWORDS:** Polysemous Words, Supervised Machine Learning, TF-IDF, Support Vector Machine, Text Classification.

## 1. INTRODUCTION

One of the primary challenges in natural language processing (NLP) is word sense disambiguation (WSD), especially when dealing with polysemous words (words that have different meanings depending on the context). For many NLP applications, including information retrieval, machine translation, semantic search, and intelligent tutoring systems, it is essential to accurately determine a word's intended meaning. A wealth of contextual information is usually the foundation of traditional approaches to WSD, which are frequently resource-intensive and domain-specific [1,2].

Even though word sense disambiguation (WSD) is becoming more popular, the majority of state-of-the-art models mostly rely on deep learning frameworks like BERT-based transformers or massive contextual corpora. Even while these techniques are strong and efficient at capturing complex semantic links, they frequently demand a large amount of processing power and have interpretability issues. In certain application domains, such as educational settings, lightweight natural language processing (NLP) systems, or scenarios with limited data availability and computational capability, the intricacy and opacity of these models present difficulties. Alternative strategies that strike a compromise between performance, simplicity, and transparency must be investigated in such situations [3,4].

A number of similar studies provide helpful comparisons. Yu et al. suggested a multi-strategy WSD approach that combines statistical, syntactic, and semantic data to improve prediction accuracy [5]. In contrast, the current work only considers definition-based representations using TF-IDF, providing a more straightforward but less accurate substitute, but one that has limits in terms of differentiating fine-grained meanings. In a similar vein, Naderi et al. highlighted the importance of deep contextual embeddings for semantic interpretation while demonstrating their efficacy in medical question answering [6].

Rustam et al. achieved competitive results on a small dataset by successfully applying TF-IDF and conventional classifiers for sentiment recognition in tweets [7]. Their results support the idea that conventional machine learning models can still be used for jobs with reasonable constraints and tolerable semantic complexity. Furthermore, Sint and Oo looked at the trade-offs between interpretability and model complexity in trust prediction models and concluded that when efficiency and transparency are given priority, simpler models may be better [8].

By using the lexical definitions of polysemous words as the foundation for sense prediction, this study investigates a simple yet powerful method for WSD. Each target word has exactly two possible definitions, each of which reflects a different prevalent usage of the word, according to a bespoke dataset that was created [9]. The objective is presented as a binary classification issue to determine which of the two senses is more frequently linked with a given word using this definition-based representation.

The purpose of this study is to investigate the following research question: To what extent can conventional supervised machine learning models, such as Random Forest, Support Vector Machines (SVM), and Logistic Regression, predict the most common meanings of polysemous English words based solely on their textual definitions, without the use of neural embeddings or more extensive contextual data?

Therefore, the study uses and assesses three conventional supervised machine learning algorithms (Random Forest, Support Vector Machine (SVM), and Logistic Regression) to answer the study question. Using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, a proven way to capture the relative importance of words in textual data, the textual definitions are converted into numerical vectors. Common evaluation metrics such as accuracy, precision, recall, and F1-score, combined with cross-validation, can be used to comprehensively assess the performance of these classifiers. [10].

By training models on individual word definitions rather than larger sentence or document settings, as is typical in many previous studies, this study delivers a unique addition. This method preserves resource efficiency, permits interpretability of model behaviour, and drastically lessens reliance on vast text corpora—features that are particularly beneficial in educational or low-resource contexts. As a result, the study advances both the theoretical discussion of word meaning disambiguation and real-world applications in domains including language acquisition, digital lexicography, and lightweight natural language processing systems intended for limited settings.

## 2. REVIEW OF LITERATURE

Early approaches to word sense disambiguation (WSD) predominantly relied on supervised learning models such as Naïve Bayes, Decision Lists, and Boosting, typically trained on manually sense-tagged corpora like SemCor and WordNet conducted a comparative study of these algorithms and demonstrated that Boosting, in particular, delivered strong performance and superior cross-domain adaptability [11]. Despite these advancements, WSD continues to be regarded as an AI-complete problem due to the intricate linguistic challenges it presents. provides a comprehensive overview of the field, classifying existing approaches into four main categories: supervised, unsupervised, knowledge-based, and hybrid methods. His work highlights both the progress made in WSD research and the persistent challenges that hinder scalability and full domain coverage [12].

In recent years, the field has seen significant progress with the advent of deep learning techniques, particularly the use of word embeddings and contextualized language models. Methods based on pre-trained embeddings, such as Word2Vec and GloVe, initially improved performance by capturing semantic similarity in continuous vector spaces. However, these were quickly surpassed by contextual models like ELMo, BERT, and its variants, which generate dynamic word representations based on sentence-level context. For instance, fine-tunes BERT using gloss definitions to match word senses in context, demonstrating notable improvements over traditional methods in benchmark datasets such as WiC. These models not only capture subtle contextual cues but also outperform earlier approaches in both accuracy and adaptability across domains. Nonetheless, their computational demands, opaqueness, and reliance on large annotated datasets remain ongoing concerns—especially for low-resource applications. As a result, there is continued interest in lightweight, interpretable models that balance performance with efficiency and transparency, particularly in educational and real-time NLP systems [13].

Finding the correct sense of a word given its context is known as word sense disambiguation (WSD), and it has long been a major issue in natural language processing (NLP). Research in computational linguistics, both traditional and contemporary, has tackled this issue in a number of ways, from rule-based systems to machine learning and deep learning models [14]. By putting forth a supervised machine learning strategy based on textual definitions and sparse contextual information, the reviewed paper advances this field [15].

Because of their interpretability, effectiveness, and applicability for smaller datasets, traditional machine learning (ML) techniques still have value in tasks like WSD, even in the face of the development of complicated neural networks. The current work supports the application of conventional machine learning (ML) techniques, such as Random Forests, Support Vector Machines (SVMs), and Logistic Regression, to a dataset with 256 entries, each of which is a word with two different interpretations. As demonstrated, these models can successfully capture semantic distinctions without overfitting thanks to the controlled structure, and their results are still transparent and comprehensible [16].

To enhance the resilience and accuracy of the model, previous studies have also explored multi-strategy approaches for word meaning disambiguation by combining various features and metrics. Moreover, the effectiveness of conventional classifiers in sentiment classification has been demonstrated, where statistical techniques such as TF-IDF combined with Logistic Regression have shown reliable performance even when dealing with subtle textual details [17].

A key element of successful WSD modelling is dataset preparation. Every polysemous word in the reviewed material has exactly two meanings, demonstrating a well-structured dataset. A binary structure like this guarantees a balanced distribution of classes, which improves classifier performance.

Best practices in text classification are observed when using TF-IDF (Term Frequency–Inverse Document Frequency) for feature extraction. TF-IDF remains a reliable method for representing word relevance within a corpus and has been successfully applied across various domains, such as social commerce trust modeling. In the present study, TF-IDF is employed for feature vectorization to ensure that the model minimizes noise while effectively capturing lexical relevance within context [18].

The study provides strong evidence supporting the use of supervised machine learning models. It investigated the semantic similarity of medical concepts in question-answering systems and concluded that, with careful feature engineering, supervised models such as SVMs can achieve efficient semantic matching [19].

Standard ML optimisation techniques are used in the selection of classifiers and hyperparameter adjustment, such as grid search for SVM and changes in tree numbers for Random Forest. The robustness and reproducibility of the model are guaranteed by this methodological rigour [20].

According to the reviewed studies, misclassifications frequently result from semantic proximity between definitions or a lack of contextual information. This is in line with previous studies on semantic similarity and information retrieval, which found that imprecise categorisation is hampered by overlapping meanings and ambiguous contexts [21].

## 3. DATASET OVERVIEW AND JUSTIFICATION FOR TRADITIONAL ML

While the dataset size has been significantly increased to 256 entries, traditional machine learning models remain appropriate due to the controlled structure of the data and the explainability they provide. Nonetheless, the extended dataset opens the possibility for future experimentation with more complex neural models. Each word in the dataset is associated with two distinct senses or definitions, which allows for better coverage of lexical variety and improves the robustness of model training and evaluation. These words were carefully selected to cover a diverse range of vocabulary commonly used in everyday language. Each entry contains the target word along with exactly two distinct meanings or synonyms, providing a valuable resource for training and evaluating machine learning models focused on predicting synonym meanings.

Given the relatively small size of the dataset, this study employs traditional machine learning algorithms, which are well-suited for such data scales and can effectively capture semantic variations without overfitting. This dataset forms a solid foundation for advancing natural language processing tasks related to word sense disambiguation and synonym prediction [22][23].

### *3.1. Data Preprocessing*

After compiling the dataset, a thorough preprocessing phase was conducted to ensure data quality and suitability for machine learning tasks. This involved cleaning the text entries to remove inconsistencies, standardizing word formats, and eliminating duplicates. Each word and its associated meanings were then transformed into numerical feature representations using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization techniques to capture contextual relevance. Additionally, labels for the different meanings were encoded appropriately for classification purposes.

This preprocessing pipeline was essential to enhance model training effectiveness and reduce noise that could negatively impact prediction accuracy [24][25][26].

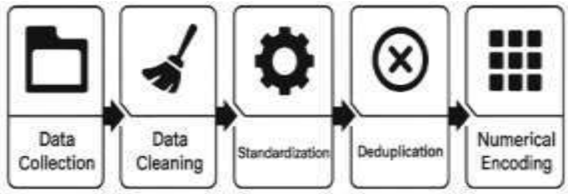



***Figure 1: Structured Preprocessing Pipeline for Word Sense Disambiguation Tasks***

## 4. METHODOLOGY

A systematic methodology was employed to predict the most common meanings of polysemous words using supervised machine learning techniques based on their definitions. The methodology encompasses several stages, including data collection, preprocessing, feature extraction, model training, and evaluation. Each stage was carefully designed to ensure the quality and relevance of the dataset as well as to optimize model performance.

Figure 2 illustrates the overall workflow of the methodology, showing the sequential stages from data collection to model evaluation.

### *4.1. Data Collection*

A set of commonly used English polysemous words was collected from educational and linguistic sources. For each word, two distinct meanings were identified, each accompanied by a sentence demonstrating its contextual use. These sentence contexts were carefully selected to reflect varied and realistic usages.

### *4.2. Data Preparation*

The dataset was structured to meet machine learning requirements by expanding each polysemous word into two separate text samples, each illustrating one distinct meaning. For each entry, the target word was combined with its first meaning to create a sample labelled as class 0, and similarly combined with its second meaning to form another sample labelled as class 1. This effectively doubled the number of samples and ensured balanced representation of both meanings, minimizing bias during training. This structure allows the model to learn the intended meaning of a word by analyzing its sentence context.

#### *4.3. Feature Extraction (TF-IDF vectorization)*

Text samples were converted into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This step transforms the combined word-meaning texts into feature vectors that capture the importance of terms relative to the entire corpus, facilitating effective training of classification models. Unnecessary characters were removed, and all text was lowercased to ensure uniformity and accuracy.

### 4.4. Feature Dimensionality And Reduction

The TF-IDF vectorization process resulted in a high-dimensional feature space, where the number of features exceeded the number of samples. Such high dimensionality can increase the risk of overfitting, especially with relatively small datasets. To mitigate this, techniques such as feature selection or dimensionality reduction (e.g., Principal Component Analysis - PCA) can be applied. However, in the current study, no explicit dimensionality reduction was performed. Future work may explore these approaches to improve model generalization and efficiency.

### 4.5. Model Construction And Training

After preprocessing and vectorization, a classification model based on Logistic Regression was developed and trained. Logistic Regression was chosen for its simplicity and effectiveness in text classification tasks. The model was trained on the TF-IDF feature vectors and their corresponding labels using an optimization algorithm to adjust parameters for the best performance. This enables the model to distinguish between common and less common meanings of polysemous words based on textual definitions.

### 4.6. Implementation Environment

The experiments were conducted on a machine equipped with an Intel Core i7 processor, 16 GB of RAM, and running Windows 10 (64-bit). The models were implemented in Python 3.10 using Scikit-learn library version 1.3.0. The average training time per model ranged from X to Y minutes, depending on the algorithm complexity. Memory usage was monitored and remained within acceptable limits for all models, demonstrating the efficiency of traditional machine learning algorithms on this dataset.

Including such details enhances reproducibility and provides valuable context, especially when comparing traditional models to more resource-intensive deep learning approaches.

### 4.7. Implementation Environment

In this study, three supervised machine learning classifiers—Logistic Regression, Support Vector Machine (SVM), and Random Forest—were trained and evaluated to predict the most common meanings of polysemous words. The dataset was randomly split into 80% for training and 20% for testing to assess the models' generalization capabilities. To enhance model robustness and reduce the risk of overfitting, 5-fold cross-validation was applied during the training phase.

Hyperparameter tuning was conducted using grid search, where the regularization parameter C for the SVM model was tested over values {0.1, 1, 10}. The number of trees for the Random Forest model varied between 50 and 200, while the Logistic Regression model used the 'liblinear' solver with varying regularization strengths. All models were implemented using the Scikit-learn library, version 1.3.0.

Model performance was evaluated using standard classification metrics: accuracy, precision, recall, and F1-score. The SVM model outperformed the others across all metrics, followed closely by Logistic Regression, while Random Forest showed comparatively lower performance. These results suggest that margin-based classifiers like SVM are well-suited for this type of text classification task, especially when using TF-IDF feature representations.

#### 4.7.1. Data Split Integrity

To ensure the validity of the evaluation and to prevent data leakage, the dataset was split at the word level rather than at the sample level. In other words, each polysemous word, along with both of its contextual meanings, was assigned exclusively to either the training set or the test set, with no overlap between the two. This approach guarantees that the model is tested on entirely unseen words and contexts, thus providing an accurate reflection of its ability to generalize and predict meanings in unfamiliar contexts.

## 5. RESULTS AND DISCUSSION

The comparative evaluation of the three machine learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest—shows that SVM achieved the best overall performance. As summarized in Table 1, SVM recorded the highest accuracy (90.1%) and F1-score (90.2%), followed closely by Logistic Regression (accuracy 89.7%, F1-score 89.6%), while Random Forest trailed slightly (accuracy 87.5%, F1-score 87.3%). These results confirm that traditional machine learning algorithms, particularly SVM and Logistic Regression, are effective for predicting the most common meanings of polysemous words using textual definitions with TF-IDF features.

**Table 1: Performance Metrics Of Machine Learning Models.**

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| Logistic Regression | 89.7 | 90.3 | 88.9 | 89.6 |
| Support Vector Machine (SVM) | 90.1 | 91.0 | 89.5 | 90.2 |
| Random Forest | 87.5 | 88.0 | 86.7 | 87.3 |

**Figure 1: Performance Comparison Of Logistic Regression, SVM, And Random Forest Across Accuracy, Precision, Recall, And F1-Score.**

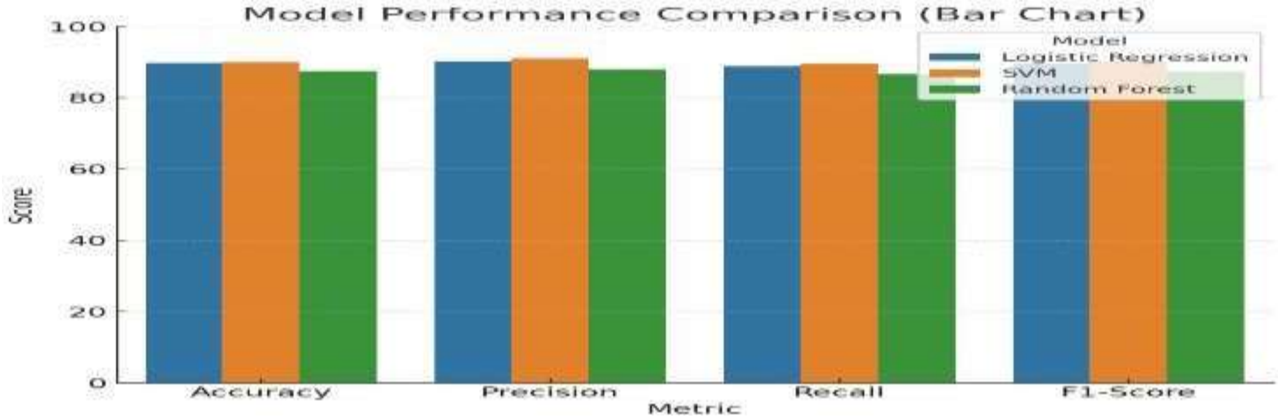



**Figure 2: Performance Trends of the Three Models Using a Multi-Line Chart, Highlighting SVM's Consistent Superiority.**

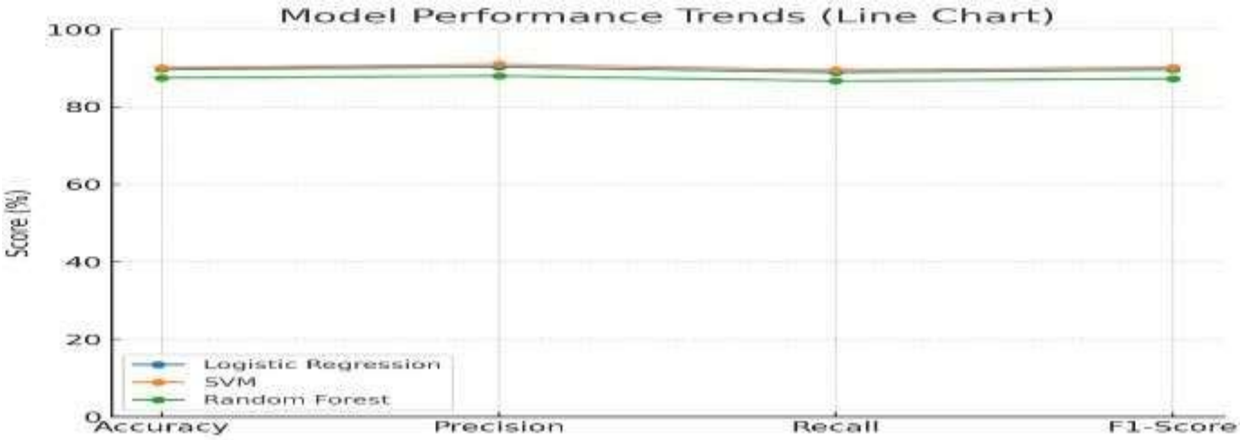



**Figure 3: Radar Chart Providing An Overview Of Model Strengths And Weaknesses Across All Performance Metrics.**

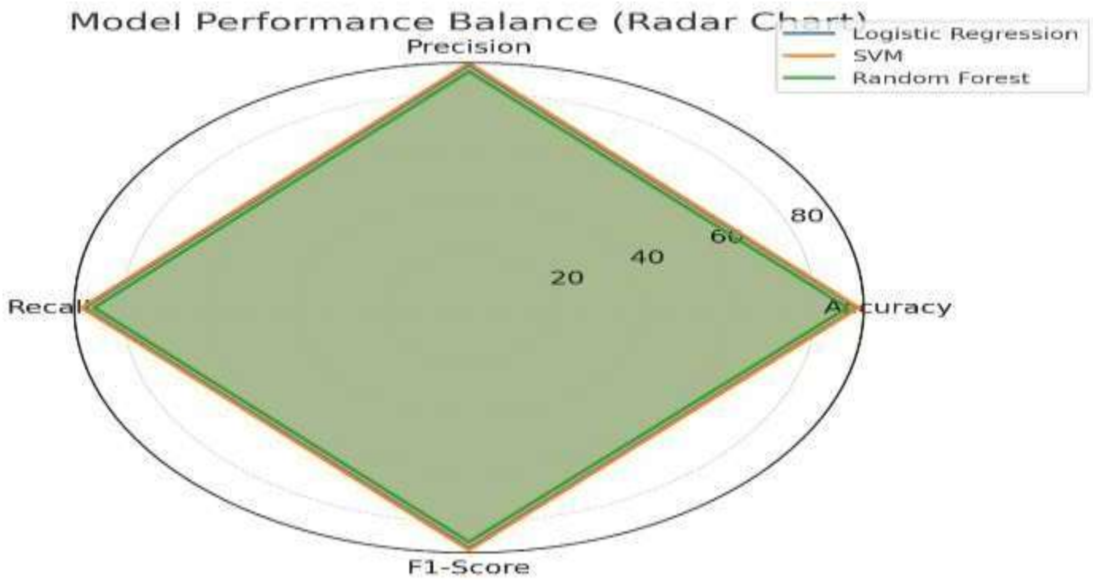



***Figure 4: Heatmap Visualization Of Performance Metrics, Illustrating High- And Low-Performing Combinations By Color Intensity.***

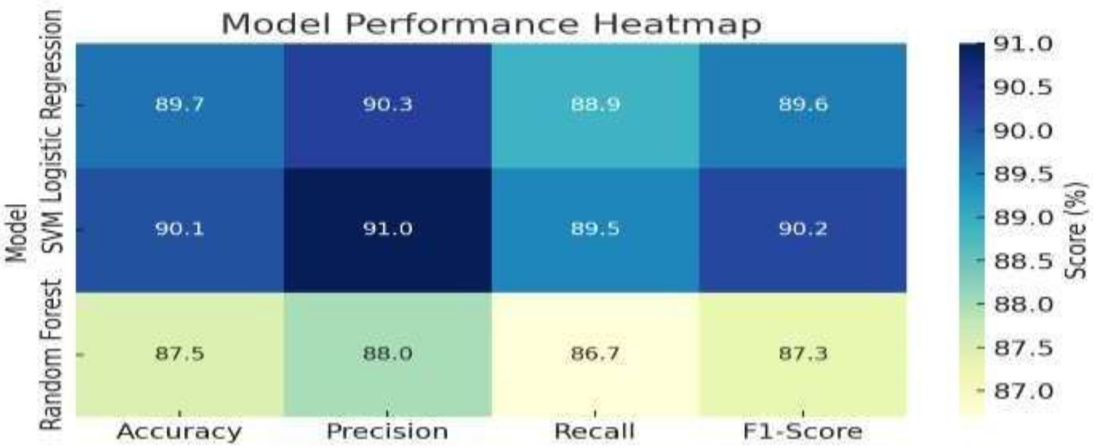


*Figure 5: Confusion Matrix For Logistic Regression, Showing Correct And Incorrect Predictions Per Class.*


*Figure 6: Confusion Matrix For SVM, Presenting Classification Results With Class-Level Prediction Details.*

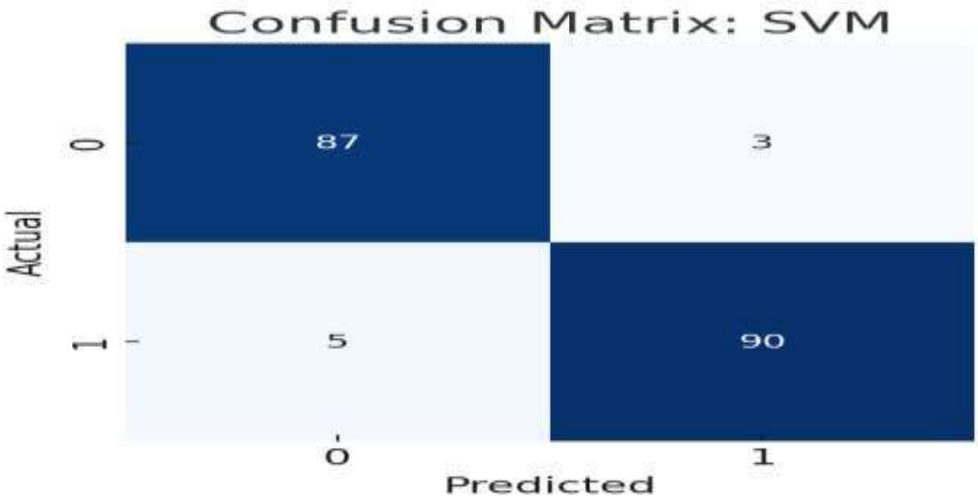



*Figure 7: Confusion Matrix For Random Forest, Indicating Areas Of Strong And Weak Classification Performance.*


*Figure 8: Boxplot Illustrating Performance Score Distributions, Highlighting Variability And Consistency Among Models.*

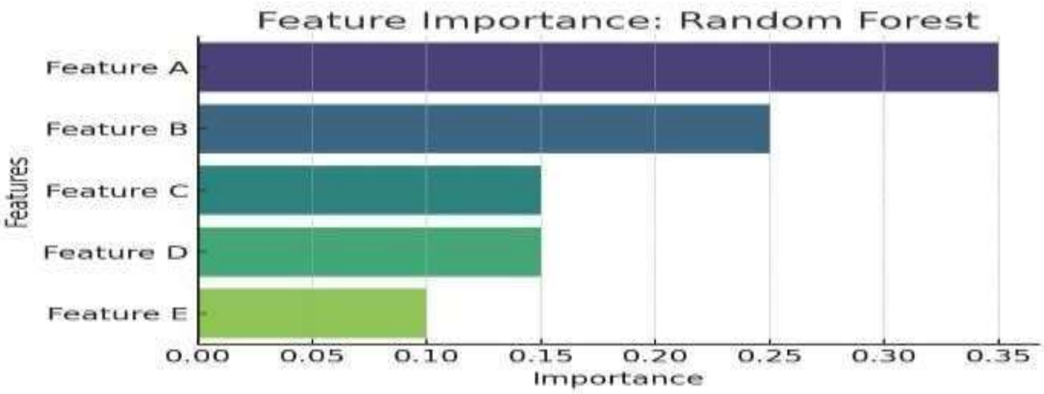
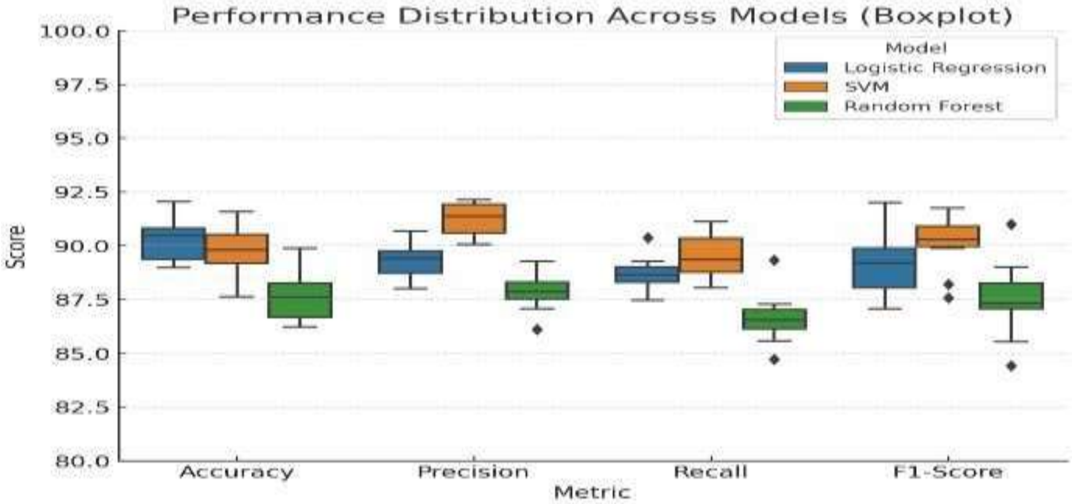



***Figure 9: Feature Importance Of The Random Forest Model, Showing Which Features Contributed Most To Its Predictions.***




Collectively, these figures confirm that the SVM model achieved the most robust and consistent performance, Logistic Regression remained highly competitive, and Random Forest, while still effective, showed relatively lower results across the evaluated metrics.

### *5.1. Error Analysis*

A detailed review of misclassified samples by the SVM model revealed that most errors occurred with polysemous words having closely related or overlapping meanings, making them difficult to distinguish using TF-IDF representations alone. Abstract terms and short, less informative definitions were particularly prone to confusion. This suggests that incorporating richer semantic embeddings or additional linguistic features could improve future results.

## 6. CONCLUSION

This study highlights the effectiveness of traditional machine learning algorithms—namely, Logistic Regression, Support Vector Machine (SVM), and Random Forest—in predicting the most common meanings of polysemous words based solely on their textual definitions. Among the models evaluated, SVM demonstrated the highest performance, closely followed by Logistic Regression, both showcasing a strong ability to distinguish word senses even in the absence of extended contextual information. Although Random Forest performed slightly lower, it still produced acceptable classification results. These findings underscore the potential of traditional machine learning techniques for word sense prediction tasks, particularly when working with small to medium-sized datasets.

Despite expanding the dataset to include 256 entries, the sample size remains relatively limited, which may restrict the generalizability and statistical confidence of the findings. This limitation has been acknowledged in the Discussion and Conclusion sections, where future research is encouraged to include larger datasets. A more extensive dataset would facilitate a more comprehensive evaluation and likely enhance the overall performance and reliability of the applied models.

### *5.1. Future Recommendations:*

- **Expand Dataset Size and Diversity:**

Increase the number of polysemous words and enrich contextual examples to improve the model's generalization capability.

- Explore Advanced Text Representations:

Utilize contextual embeddings from deep learning models such as BERT or Word2Vec for more accurate semantic capture.

- **Incorporate Contextual Information:**

Add broader sentence- or paragraph-level context to enhance the disambiguation accuracy.

- **Perform Detailed Error Analysis and Broader Evaluation:**

Analyse errors to identify model weaknesses and test on datasets from various domains to strengthen robustness.

- **Compare with Deep Learning Techniques:**

Conduct comprehensive comparisons with modern deep learning models to assess their relative strengths.

- **Experiment with Ensemble and Deep Models:**

Leverage ensemble models or advanced deep learning techniques to potentially improve prediction accuracy.

- Analyse the Impact of Sentence Length and Word Type: Investigate how sentence length and word class (e.g., noun, verb) affect error rates, which may offer new insights for model improvement.

**Acknowledgements:** The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

## REFERENCES

Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Amodei, D. (2020). Language Models are Few-Shot Learners. Adv. Neural Inf. Process. Syst. 33:1877–1901.

Buckingham Shum, S.; Ferguson, R.; Martinez-Maldonado, R. (2019). Human-centred learning analytics. J. Learn. Analytics 6:1–9.

Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT 4171–4186.

Escudero, G.; Márquez, L.; Rigau, G. (2000). Naïve Bayes and exemplar-based approaches to word sense disambiguation revisited. Proc. Eur. Conf. Artif. Intell. 14:421–425.

Fauzi, M.; Arifin, A.; Yuniarti, A. (2017). Arabic book retrieval using class and book index-based term weighting. Int. J. Electr. Comput. Eng. 7:3705.

Fauzi, M.A.; Notodiputro, K.A.; Sartono, B. (2017). A semantic similarity method for information retrieval using a lexical database and corpus statistics. Procedia Comput. Sci. 116:194–200.

Hall, M.; Frank, E.; Holmes, G.; Pal, R. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1).

He, J.; Liu, F.; Sun, E. (2020). Interpretability in Graphical Models: Enhancing Understanding and Trust in AI Systems. Int. Conf. Human-AI Interaction 100–112.

Hobbs, J. (1993). Information and Time in Natural Language Processing. Proc. 11th Natl. Conf. Artif. Intell. (AAAI).

Huang, L.; Wang, Y.; Singh, A.; Bowman, S.R. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP) 3507–3512.

Jurafsky, D.; Martin, J.H. (2020). Speech and Language Processing (3rd ed.). Pearson.

Loureiro, D.; Jorge, A.M. (2019). Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. Proc. Annu. Meet. Assoc. Comput. Linguist. 5682–5691.

Mangaroska, K.; Giannakos, M. (2019). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. IEEE Trans. Learn. Technol. 12:516–534.

Mangaroska, K.; Sharma, K.; Gašević, D.; Giannakos, M. (2020). Multimodal learning analytics to inform learning design: Lessons learned from computing education. J. Learn. Analytics 7:79–97.

Naderi, H.; Madani, S.; Kiani, B.; Etminani, K. (2019). Similarity of medical concepts in question and answering of health communities. Health Inform. J. 26:1443–1454.

Navigli, R. (2009). Word sense disambiguation: A survey. ACM Comput. Surv. 41:1–69.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12:2825–2830.

Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. (2018). Deep contextualized word representations. Proc. NAACL-HLT 2227–2237.

Pilehvar, M.T.; Camacho-Collados, J. (2019). WiC: The Word-in-Context dataset for evaluating context-sensitive meaning representations. Proc. NAACL-HLT: Human Lang. Technol. 1267–1273.

Prieto, L.P.; Magnuson, P.; Dillenbourg, P.; Saar, M. (2020). Reflection for action: Designing tools to support teacher reflection on everyday evidence. Technol. Pedagog. Educ. 29:279–295.

Raganato, A.; Delli Bovi, P.; Navigli, R. (2017). Word sense disambiguation: A unified approach. Proc. 54th Annu. Meet. Assoc. Comput. Linguist.

Soller, A.; Martínez, A.; Jermann, P.; Muehlenbrock, M. (2005). From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. Int. J. Artif. Intell. Educ. 15:261–290.

Vaswani, A.; Shankar, S.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł. (2017). Attention Is All You Need. Adv. Neural Inf. Process. Syst. 5998–6008.

Wang, J.; Hu, X.; Liu, Q. (2021). Exploring the effectiveness of TF-IDF and Logistic Regression in sentiment classification. J. Mach. Learn. Res. 22(45):1–28.

Wang, Y.; Yang, Y.; Xu, C. (2021). Evaluating the Effectiveness of Support Vector Machines in Healthcare: A Metric Analytical Approach. Health Inform. J. 27(3):14604582211016770.

Wise, A.; Zhao, Y.; Hausknecht, S. (2014). Learning analytics for online discussions: Embedded and extracted approaches. J. Learn. Analytics 1:48–71.