Journal of Information Systems Engineering and Management
2025, 10(9s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Context-Aware Transformer Models for Ambiguous Word Classification in Code-Mixed Sentiment Analysis

Shruti Mathur[1], Dr Gourav Shrivastava[2]
[1] Research Scholar, Sanjeev Agrawal Global Educational University, Bhopal, Madhya Pradesh, India
[2] Professor, Sanjeev Agrawal Global Educational University, Bhopal, Madhya Pradesh, India
shrutimathur19@gmail.com

ARTICLE INFO

Received: 30 Oct 2024

Revised: 19 Dec 2024

Accepted: 02 Jan 2025

ABSTRACT

Introduction: In text-based sentiment analysis, an ambiguous word that has more than one meaning can result in ambiguity, which creates challenging issues in analysing the sentiment. The Deep Learning models have achieved effective classification of ambiguous words. The traditional models acquirecertain limitations in accuracy and efficiency due to the ignorance of context features and parallelization.

Objectives: To classify ambiguous words in code-mixed sentiment analysis using context-aware transformer models, improving accuracy, efficiency, and adaptability while addressing limitations in traditional methods.

Methods: This study employs transformer-based models (DistilBERT, IndicBERT, XLM-RoBERTa, TinyBERT) to classify ambiguous words in code-mixed text. Data preprocessing techniques, including stopword removal, stemming, and lemmatization, prepare inputs for training. Each model's performance is evaluated using metrics like accuracy, precision, recall, and F1-score. DistilBERT's superior efficiency is compared against others to identify the best-suited approach for ambiguous word classification.

Results: The study demonstrates the effectiveness of transformer models for context-aware ambiguous word classification in code-mixed sentiment analysis. DistilBERT outperforms others with an accuracy of 88.75%, precision of 92.87%, recall of 88.75%, and F1-score of 90.39%. Its lightweight architecture ensures faster inference and reduced memory usage compared to IndicBERT, TinyBERT, and XLM-RoBERTa.

Conclusions: The findings confirm DistilBERT's robustness and reliability for code-mixed language tasks, offering significant advancements over conventional methods by achieving superior accuracy and efficiency in sentiment analysis. Context-aware transformer models are uniquely optimized for ambiguous word classification in code-mixed sentiment analysis, ensuring high precision, efficiency, and applicability to multilingual challenges.

Keywords: Ambiguous word classification, Bidirectional Encoder Representations from Transformers, sentiment analysis, multi-head attention, transformer models

1. INTRODUCTION

Social networks such as Twitter, Facebook, and various social media platforms produce vast amounts of short text, where the classification of textbased on sentiment is the main hotspot in Natural Language Processing(NLP). Most of the sentences in English have different forms of ambiguitythatinclude anaphoric, constituent boundary, homograph, syntactic, semantic, and internal word structure, which results indifficultiesclassifying the word ambiguous in sentiment.The word “Duck” can mean "a bird" or "bend"and is a simple example of ambiguity [1][2][3].Word sense disambiguation (WSD) is a word classification task, where the ambiguous word in the sentence is split into single words using WSD-based models.InNLP, there are various polysemous words to determine the correct ambiguous word in the sense of context, where disambiguation is considered a challenging task in Deep Learning|(DL) [4] [5]. To overcome these challenges Machine Learning (ML) techniques are introduced for efficient word classification using algorithms such as Naive Bayes [6], decision tree (DT) [7], and support vector machines (SVMs) [8].

In recent days, various techniques were introduced for ambiguous word classification but the models were limited to capturingthe complex relationship among the words, however, the DL techniques which involve convolutional neural networks (CNNs) [9] and recurrent neural networks (RNNs) learned from the raw text data. The model directly converts the data into a numerical format and builds embedding, recurrent, and dense layers, which are compiled with suitable loss functions and optimizers to higher the accuracy in word classification [10]. These modelsutilized implicit and explicit representation concepts, where the implicit collected all data from the dictionary except the data with sparsity, on the other hand, the explicit collected the semantic data, through which the sparsity and ambiguity problems are solved with higher accuracy in word classification [2]. Graph Neural Network (GNN)[11] was one of the DL techniques that outperformed well in the classification of ambiguity in sentiment analysis.The sequential learning models directly operated GNN on the graph structure and captured the sequential information from the graph, while the model faced challenges due to a complex relationship between the nodes [10]. The dynamic CNN (DCNN) [12] used max-pooling and wide convolutionaltechniques, which reduce the risk of outfitting, and dimensionality, and achieve higher performance in classification, however, the model was limited to the amount of information in the context [4].

The research analyzes various transformer-based modelsincluding BERT, Tiny BERT, XLM-Roberta, and Indic BERT for efficient ambiguousword classification. These models easily understand the contextually represented words and show adaptability in fine-tuning the classification models. Each model is trained and their performance is evaluated, where the best model is selected based on performance metrics.

The research is detailed in the following section as follows. Section 2 describes the literature review with advantages and disadvantages. Section 3 explains the proposed methodology; section 4 describes the experimental results, comparative analysis, performance metrics, and discussion. Section 5describes the conclusion with future work.

2. LITERATURE REVIEW

Katherine A. De Long *et al.* [13] deployed an ambiguous word classification model using the BERT pre-trained neural language, where a continuous rating method reliant on the linguistic phenomenon of zeugma evaluated word factor beyond the dictionary and showed accurate enhancement word classification, however, the model was limited and dictionary-based categories provided with BERT. Sanaa Kaddoura *et al.* [1] presented ambiguous word classification using the BERTpre-trained language model, which involves a weighted voting model that maximizes the weight to enhance the word classification process and the BERT model outperformed the benchmark algorithm.Even though the model showed higher accuracy, the generalization concept caused the ineffectiveness while classification. Chun-Xiang Zhang *et al.* [4] designed a model for word sense disambiguation that ensemble multi-head self-attention and gated-dilated convolution mechanism. The adaptive average pooling layer and multi-headed self-attention were adapted to compute the ambiguous word weight and learn the connection among discriminative features with higher accuracy and efficiency in the word classification process. However,the model showed higher effectiveness,while the multi-head self-attention model was limited to identifying the ambiguity. AytugOnan[10]developed a graph-based word classification using BERT-based dynamic fusion, which identified the connection among the nodes and hierarchical graph with accuracy.The developed model was only limited tothe English language. Yingying Liu *et al.* [2] introduced a short text classification process using the CNN and Temporal Convolutional Network (TCN) ensemble model, where performance was effective and efficient in short word classification yet the model failed to obtain the character-level semantics information.

3. METHODOLOGY OF TRANSFORMER-BASED MODELS FORAMBIGUOUS WORD CLASSIFICATION

The research analyzes different transformer-based models to classify the ambiguous word, where the input data is collected from the real-time dataset and the data is given to the pre-processing technique, which involves special character removal, stop word removal, lemmatization, and stemming, where the raw data is processed for modelling, and analyzing. Pre-processed words are further trained using some of the training models involved such as XLM-Roberta, Indic BERT, Distil BERT, and Tiny BERT for the classification of words based on past observation. At last, the trained model classifies the ambiguous word with high accuracy and performance whereas the Digital BERT model outperforms other models.Figure 1 illustrates the flow diagram of ambiguous word classification.

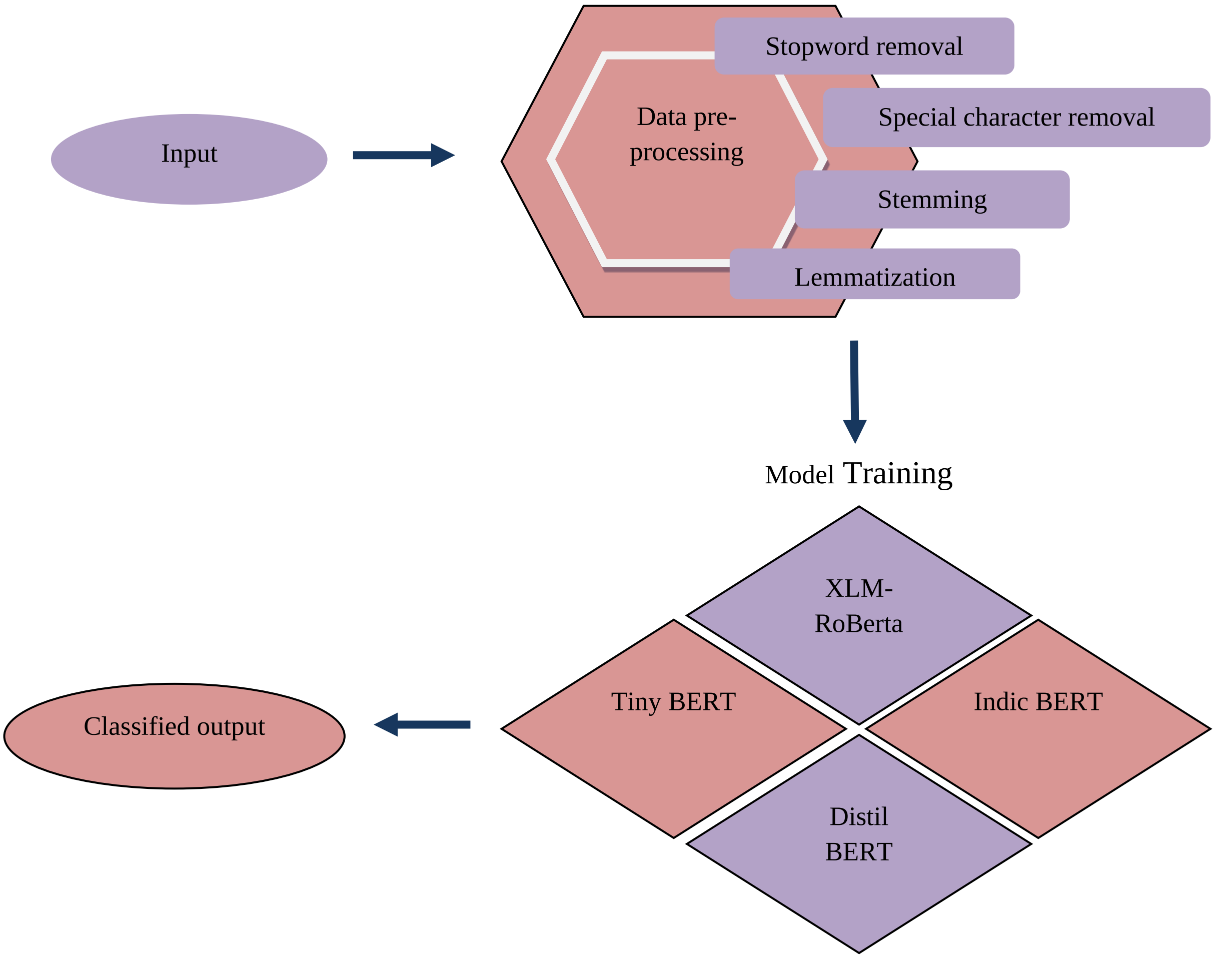


Figure 1 Ambiguous word classification workflow

3.1 Input text

The input text is taken from the real-time dataset, which is in the form of structured data or tokens and split into "N" numbers of words for efficient classification.The input text dataset is mathematically represented as

$$I \ =\{I_1, I_2,...,I_i,...,I_N\} \tag{1}$$

where, I_i represents i^{th} the word in the dataset (I) and I_N term represents the total number of words in the dataset.

3.2 Text preprocessing

The input text is preprocessed usingthe textpreprocessing techniques, where frequently appearing words are eliminated using the stop word removal technique, special character removal uses regular expressions or string manipulation functions to replace the special character into an empty string. In line with this, the input data are reduced to a common base root using stemming and lemmatization techniques,enhancing the usability and quality of words for successive modelling and analysis. The preprocessed word is represented below.

$$I^* = \{I_1^*, I_2^*.......I_i^*......I_N^*\} \tag{2}$$

where $\left(I^*\right)$ denotes the preprocessed dataset, I_i^* denotes the i^{th} preprocessed text, and I_N^* denotes the total number of preprocessedtexts.

3.3 Models for Ambiguous Word Classification

The research utilizes transformer-based models for training the pre-processed data which include, XLM-Roberta, Indic BERT, Distil BERT, and Tiny BERT are briefly described withtheir workflow in the following section.

3.3.1 XLM-Roberta for ambiguous word classification

In the XLM Roberta model, the preprocessed input data is in the form of encoded data, where the model leverages the pre-trained features and obtains a raw classification score for each category of word classification. The softmax function is applied to convert the logits into probabilities representation and the classification is done by decision, which helps in the effective classification of ambiguous words [14].

3.3.2 Indic BERT for ambiguous word classification

The Indic BERT utilizes 6 hidden encoders and decoders with filter sizes of 1024 and 4096 respectively for word classification. According to Poisson distribution, $(\lambda = 3.5)$ the sentence is masked randomly by sampling span length. The Indic BERT model utilises an Adam optimizer with a higher learning rate of 0.0001 that sped up the convergence and resulted in quality outcome,batch size of 4096 tokens, and label smoothing of 0.1 for effective training in word classification. Then the pre-trained model is compressed by training the model with cross-layer parameters and to overcome the single script representation, the model is again trained with 64K vocabulary using the original script. The IndicBERT model shows higher performance for low-resource languages [15].

3.3.3 Distil BERT for ambiguous word classification

The Distil BERT model takes a preprocessed word (I^*) as an input, which is converted into a group of embedding words. Each word in the contextual embedding is combined into a unified word, which denotes the correct meaning of the original sentence then it is passed into the fully connected layer, which outputs a vector size $(I^*{}_N)$.The Distil BERT model focuses on three main objectives for effective classification Distillation loss, masked language modeling (MLM), and Cosine embedding loss. In Distillation loss, the model is optimized to equalize the base model for equal probability in word classification. The MLM includes the random masking of 15% of preprocessed words in sentences. Unlike other autoregressive models like Generative Pre-Trained Transformers (GPT) or convolutional models like RNN, the distil BERT marks the feature token with an internal process and obtains sentence representation in a bi-directional process. Figure 2 represents the components and architecture of distil BERT in ambiguous word classification [16].

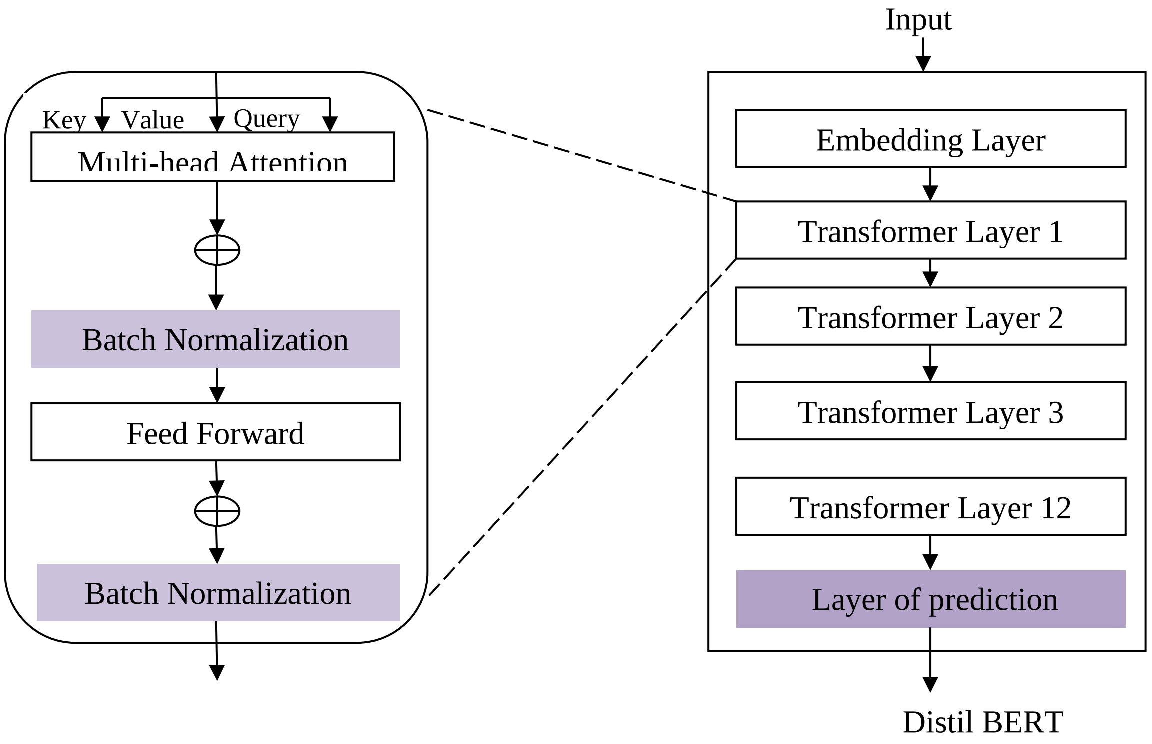


Figure 2 Component and architecture of Distil BERT

3.3.4 Tiny BERT for ambiguous word classification

The Tiny BERT model ensures the task-specific and general domain in the BERT with reduced model size and inference time. The model occupies four layers with 312 hidden sizes, 1200 feed-filter sizes, and 12 head numbers that have 14.5M parameters. The model utilizes a $3*m$ layer mapping function and one learning weight for each layer. The general distillation and the task-specific distillation are involved in Tiny BERT learning, where the general distillation sets 128 as a sequence length and performs word classification in intermediate layer distillation for 3 epochs whereas, in the task-specific distillation, there are 20 epochs with learning rate and base size of 32, the model performs classification for 3 epochs by choosing learning rate from $\{1e-5\}$ and the batch size from $\{16,32\}$. The Tiny BERT model set 64 as a sequence length for single sentence tasks and 128 for pair sequence tasks for word classification [17].

4. RESULT

The result section describes the experimental setup, performance metrics, and comparative analysis with other existing training models.

4.1 Experimental setup

The ambiguous word classificationexperiment is executed on the Windows 11 operating system (OS) using PyCharm software with 128 GB ROM and 16GB RAM.

4.2 Performance metrics

The performance metrics used in the effective analysis such as Precision, Accuracy, F1-score, and Recall are described below with their mathematical representation.

A) Accuracy

Accuracy metric is defined as the ratio of correct classification outcome to the total number of outcomesand is mathematically denoted as,

$$Accuracy = \frac{TP_w + TN_w}{TP_w + TN_w + FP_w + FN_w} \tag{1}$$

B) Precision

The precision measures the positive instances during classification and ismathematically represented as,

$$\mathrm{Pr}ecision = \frac{TP_w}{TP_w + FP_w} \tag{2}$$

C) F1-score

The f-1 score computes the average ofrecall and precision, whichis mathematically denoted as,

$$F1 - score = 2 * \frac{precision * \mathrm{Re}call}{precision + \mathrm{Re}call} \tag{3}$$

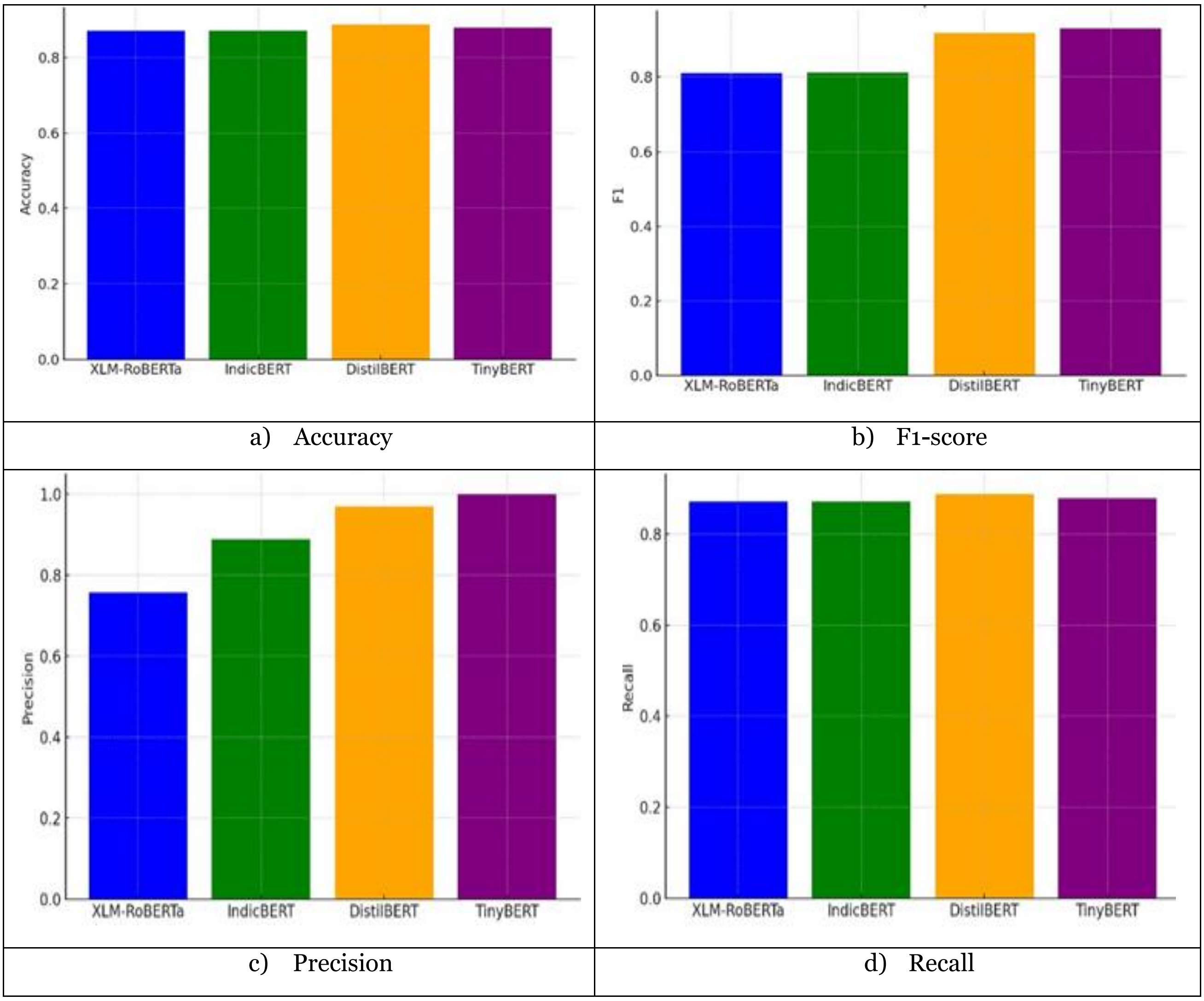
D) Recall: Recall measures the ratio of the true positive to the total of true positive and false negative,which is mathematically denoted as

$$\mathrm{Re}call = \frac{TP_w}{TP_w + FN_w} \tag{4}$$

where TP_w denotes the true positive, TN_w denotes the true negative, FP_w denotes the false positive, and FN_w denotes the false negative.

4.3 Comparative Analysis

In this section, various Transformer-based models such as XLM-Roberta [14], Indic BERT[15], Distil BERT[16], and Tiny BERT [17]are analyzed based on their performance for ambiguous word classification, where the model XLM-Roberta shows higher recall metrics however it is limited to Hinglish adaptation with lower precision rate. The Indic BERT model obtained a higher recall and precision rate but was poor in the F1-score. The Tiny BERT model acquired maximal F1-score with a lower recall rate comparing these three models the Distil BERT model outperformed with higher and stable performance of balanced metrics for the word classification. Figure 3 represents the comparative analysis for ambiguous word classification.



a) Accuracy

b) F1-score

c) Precision

d) Recall

Figure 3: Comparative analyses of different models in ambiguous word classification

4.4 Comparative discussion

In this section, each model is discussed with its performance metrics, where the XLM-RoBERTa is trained with multilingual corpora and effective on high recall metrics but shows moderate precision with an accuracy of 87.04%.The IndicBERT shows a good precision rate with a lower accuracyof 87.14% whereas the DistilBERTresults in balanced performance witha higher accuracy of 88.75%in ambiguous word classification. The Tiny BERT showsa higherprecision rate but struggles with recall and consistency with an accuracy of 87.85%. Moreover, some of the traditional techniques including WSD [4], CRFA [2], BERT [18], and STCKA [19]attain accuracy of 73.86%, 80.96%,43.85%, and79.3%. Analyzing recent models,where the DistilBERT works with higher accuracy and is easily adapted to efficient classification. Table 1 shows the comparative discussion of different models with their performance metrics.

Table 1 Comparative discussion for ambiguous word classification

Models vs. Metrics	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
AGDCNN-based WSD	73.86	77.42	76.84	77.12
CRFA	80.96	77.62	77.63	77.57
BERT	43.85	42.79	43.85	37.66
STCKA	79.3	76.79	76.54	76.63
XLM-RoBERTa	87.04	75.77	87.04	81.02
IndicBERT	87.14	88.8	87.14	81.15
TinyBERT	87.85	100	87.15	93.13
DistilBERT	88.75	92.87	88.75	90.39

5. CONCLUSION

Ambiguous word classification in sentiment analysis is the most significant task in NLP, where several transformer-based techniques and approaches are utilized for effective word classification. The research focuses on analyzing different transformer-based modelsfor ambiguous word classification including XLM-Roberta, Indic BERT, Distil BERT, and Tiny BERT, where the Distil BERT model outperforms the three models by focusing on some of the objectives including distillation loss, MLM, Cosine embedding loss.The Distil BERT model is a light and smaller transformer, which obtains the best performance with minimum computational time and spacein their classification. The efficiency of the Distil BERT model is analyzed with precision, recall, F1-score metrics accuracy, and recall, which attain a maximum of92.87%, 88.75%, 90.39%, 88.75%, and 88.75%, respectively. The overall analysis shows that the Distil BERT model provided a significant outcome in ambiguous word classification. In the future,the Distil BERT model will be combined with DL models to improve the accuracyof sentiment analysis without ambiguity issues.

REFERENCE

[1] Kaddoura S, Nassar R. EnhancedBERT: A feature-rich ensemble model for Arabic word sense disambiguation with statistical analysis and optimized data collection. Journal of King Saud University-Computer and Information Sciences. 2024 Jan 1;36(1):101911, https://doi.org/10.1016/j.jksuci.2023.101911.

[2] Liu Y, Li P, Hu X. Combining context-relevant features with multi-stage attention network for short text classification. Computer Speech & Language. 2022 Jan 1;71:101268, https://doi.org/10.1016/j.csl.2021.101268.

[3] Abou-Khalil, V., Helou, S., Flanagan, B. et al. Learning isolated polysemous words: identifying the intended meaning of language learners in informal ubiquitous language learning environments. Smart Learn. Environ. 6, 13 (2019), https://doi.org/10.1186/s40561-019-0095-0.

[4] C. -X. Zhang, Y. -L. Zhang and X. -Y. Gao, "Multi-Head Self-Attention Gated-Dilated Convolutional Neural Network for Word Sense Disambiguation," in *IEEE Access*, vol. 11, pp. 14202-14210, 2023, doi: 10.1109/ACCESS.2023.3243574.

[5] Jaber A, Martínez P. Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques. Methods of Information in Medicine. 2022 Jun;61(S 01):e28-34, DOI: 10.1055/s-0042-1742388.

[6] A. Abraham *et al.*, "Naïve Bayes Approach for Word Sense Disambiguation System With a Focus on Parts-of-Speech Ambiguity Resolution," in IEEE Access, vol. 12, pp. 126668-126678, 2024, doi: 10.1109/ACCESS.2024.3453912.

[7] H. Liu, P. Burnap, W. Alorainy and M. L. Williams, "A Fuzzy Approach to Text Classification With Two-Stage Training for Ambiguous Instances," in IEEE Transactions on Computational Social Systems, vol. 6, no. 2, pp. 227-240, April 2019, doi: 10.1109/TCSS.2019.2892037.

[8] D. W. Otter, J. R. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 2, pp. 604-624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.

[9] Yepes AJ. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. Journal of biomedical informatics. 2017 Sep 1;73:137-47, https://doi.org/10.1016/j.jbi.2017.08.001.

[10] Onan A. Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion. Journal of king saud university-computer and information sciences. 2023 Jul 1;35(7):101610, https://doi.org/10.1016/j.jksuci.2023.101610.

[11] Zhang CX, Liu R, Gao XY, Yu B. Graph convolutional network for word sense disambiguation. Discrete Dynamics in Nature and Society. 2021;2021(1):2822126, https://doi.org/10.1155/2021/2822126.

[12] Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications. 2017 Apr 15;72:221-30, https://doi.org/10.1016/j.eswa.2016.10.065.

[13] DeLong, K.A., Trott, S. & Kutas, M. Offline dominance and zeugmatic similarity normings of variably ambiguous words assessed against a neural language model (BERT). *Behav Res* **55**, 1537–1557 (2023), https://doi.org/10.3758/s13428-022-01869-6.

[14] Gaurav A, Gupta BB, Sharma S, Bansal R, Chui KT. XLM-RoBERTa Based Sentiment Analysis of Tweets on Metaverse and 6G. Procedia Computer Science. 2024 Jan 1;238:902-7, https://doi.org/10.1016/j.procs.2024.06.110.

[15] Dabre R, Shrotriya H, Kunchukuttan A, Puduppully R, Khapra MM, Kumar P. IndicBART: A pre-trained model for indic natural language generation. arXivpreprint arXiv:2109.02903. 2021 Sep 7, https://doi.org/10.48550/arXiv.2109.02903.

[16] Nair AR, Singh RP, Gupta D, Kumar P. Evaluating the Impact of Text Data Augmentation on Text Classification Tasks using DistilBERT. Procedia Computer Science. 2024 Jan 1;235:102-11. https://doi.org/10.1016/j.procs.2024.04.013.

[17] Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, Liu Q. Tinybert: Distilling bert for natural language understanding. arXivpreprint arXiv:1909.10351. 2019 Sep 23, https://doi.org/10.48550/arXiv.1909.10351.

[18] Kenton JD, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. InProceedings of naacL-HLT 2019 Jun 2 (Vol.1, p.2), https://doi.org/10.18653/v1/N19-1423.

[19] Chen J, Hu Y, Liu J, Xiao Y, Jiang H. Deep short text classification with knowledge powered attention. InProceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 6252-6259), https://doi.org/10.1609/aaai.v33i01.33016252.