WSD-AF: Word Sense Disambiguation and Adaptive Fusion for Text Classification

Mandy Huang, Fengfei Yu, and Danlin Luo
McGill University

Abstract

Text classification has witnessed significant advancements in recent years, largely driven by deep neural network models. However, our investigation reveals that current DNN-based classification approaches struggle to handle noisy sentence inputs, especially those with implicit negation, self-conflicting semantics, or other complex linguistic structures. To address these challenges, we propose WSD-AF, a novel framework that combines word sense disambiguation with adaptive fusion to enhance semantic richness and reduce internal data noise. While experimental results show that WSD-AF achieves performance comparable to baseline models in accuracy and F1 score, it potentially has improved robustness in handling noisy and adversarial inputs, highlighting its potential for real-world applications requiring fine-grained text classification.

1 Introduction

Sentiment analysis is a natural language processing technique that derives the emotions expressed in a piece of text. The value of sentiment analysis lies in the ability to gain qualitative insights on large corpora of data. One example of the real-life usage of sentiment analysis is in product marketing. Companies can use the results of sentiment analysis to improve their products or modify their marketing techniques (Birjali et al., 2021). The domain of analysis may also vary depending on the target audience, allowing for a wide variety of applications in sentiment analysis.

However, sentiment analysis is often tricky due to ambiguity and noisiness that arise from language. Traditional sentiment analysis struggles when entities in a text are ambiguous or context-dependent, such as when the sentiment associated with "Apple" refers to either the company or the fruit depending on the context. The multiple meanings of words can create ambiguity for the machine. Additionally, certain types of text, such as tweets,

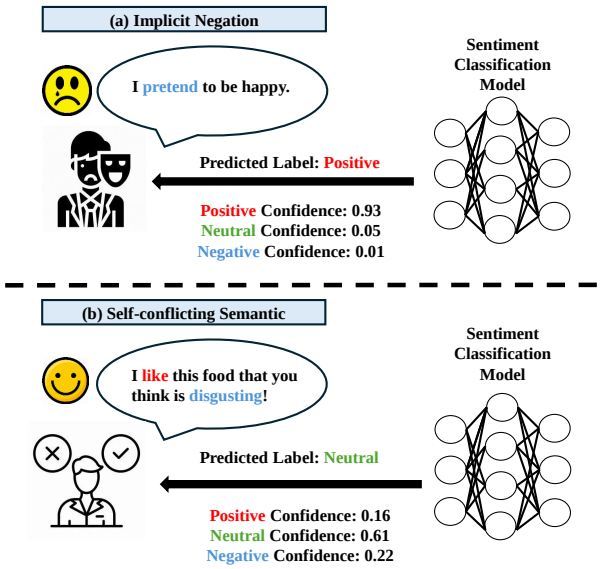

Figure 1: Motivation: DNN-based classification model cannot (a) capture implicit semantics; (b) handle noisy inputs with self-conflicting semantics.

differ greatly from traditional language and cause noisiness (Meisheri et al., 2017). Such challenges highlight the need for approaches that can improve sentiment analysis by accurately identifying entities and, most importantly, understanding their specific context. This project hypothesizes that the integration of entity recognition and entity disambiguation can significantly enhance sentiment analysis by addressing these issues.

To address challenges such as conflicting semantics, ambiguity, and complicated linguistics, we propose WSD-AF, a novel framework that combines word sense disambiguation with adaptive fusion to enhance semantic richness and reduce internal data noise. Our framework first applies WSD to extract fine-grained semantic representations, followed by an adaptive fusion mechanism that dynamically integrates the contextual vector (local semantics) and global vector (overall seman-

tics). This enables the model to better capture the core meaning of text while mitigating the impact of noisy or conflicting signals. Additionally, Word Sense Disambiguation (WSD) has long been a challenge in natural language processing, particularly when handling multi-word expressions like "Latin America" and reconciling tokenization mismatches with lexical resources. With the advent of contextualized embeddings like BERT, this study adapts embeddings to process phrases as single entities and resolves tokenization issues to improve WSD accuracy.

While experimental results show that WSD-AF achieves performance comparable to baseline models in terms of accuracy and F1 score, it demonstrates potential advantages in robustness when dealing with challenging text inputs. These findings highlight the framework's applicability in real-world tasks that demand fine-grained text classification under noisy conditions. Sec. 2 discusses related works, and Sec. 3 delves into the motivation and specific challenges addressed by the WSD-AF framework. We explore the structure of the framework in Sec. 4. In Sec. 5, the setup of the experiment is outlined. Finally, we discuss the results in Sec. 6 and explore the limitations of this study in Sec. 7.

2 Related Work

Although text classification has been a cornerstone of NLP, the robustness of existing models under noisy or ambiguous input conditions remains a persistent challenge. Several approaches have been proposed to address this issue, focusing on label noise, semantic ambiguity, and adaptive learning.

Previous research has explored methods to mitigate the effects of noisy labels in training data. Techniques such as noise-robust training frameworks (Song et al., 2022) and adversarial training methods (Miyato et al., 2021) have demonstrated the ability to improve model performance under noisy conditions. These methods typically rely on refining training objectives or dynamically adjusting model learning to avoid overfitting to noisy labels. However, they often overlook the inherent complexity in text inputs, such as implicit negation or self-conflicting semantics. Lee et al. (2022) has demonstrated that adversarial training at the context level improves model performance against noise, indicating that addressing noise as our WSD-AF framework does will allow the model to be resilient against a variety of inputs.

Word sense disambiguation (WSD) has been a critical area in improving NLP applications, enabling models to resolve lexical ambiguity for better semantic understanding. Huang et al. (2019) introduced GlossBERT, a method that fine-tunes the BERT transformer model with gloss knowledge to achieve state-of-the-art results in WSD tasks. Research has demonstrated the importance of disambiguation techniques by integrating WSD into classification pipelines (Pilehvar et al., 2017). Although the article also explores the effect of creating a pipeline with disambiguation on model performance, our pipeline specifically targets sentiment analysis against noisy inputs.

3 Motivation

In real-world scenarios, large-scale data collection often introduces significant noise in textual inputs. Sentences may vary in expression depending on the speaker, leading to discrepancies in the fine-grained sentiment conveyed within the text. This challenge is particularly relevant in real-life applications, such as understanding the emotional expressions of individuals with autism. These individuals often do not explicitly articulate their emotions; instead, they use implicit language to convey their inner thoughts. In some cases, they might even employ contradictory phrases to obscure their true feelings, presenting a unique challenge for sentiment analysis.

To explore these challenges, we conducted simple motivational experiments. As shown in Fig. 1 (a), for sentences with **implicit negation** (e.g., "pretend to do something"), current models fail to classify the sentiment accurately. In Fig. 1 (b), models struggle to identify the dominant sentiment within sentences containing conflicting emotions. Instead, they are misled by word frequency or sentiment intensity, which highlights the **self-conflicting sentiment/semantic** nature of such adversarial inputs. These findings reveal a critical limitation in existing models: their inability to effectively handle noisy inputs and capture nuanced emotional expressions.

This observation raises an important question: *How can a model capture fine-grained emotions and accurately classify sentences with noisy and adversarial inputs?* To address this, we propose the **WSD-AF framework**, which integrates **Word Sense Disambiguation (WSD)** to enhance seman-

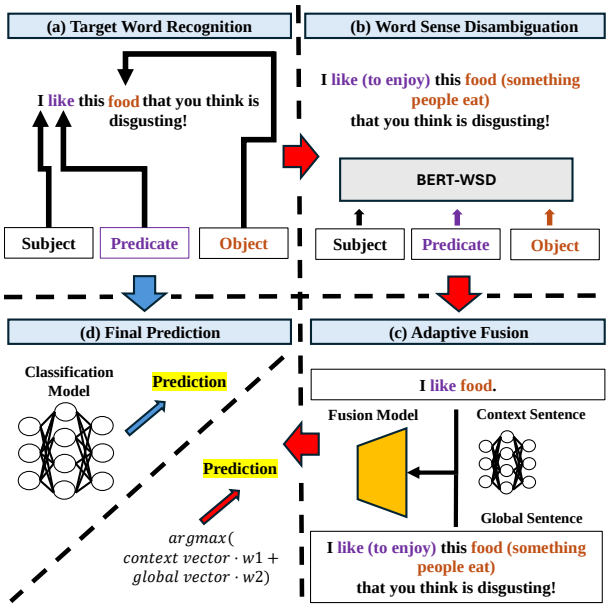

Figure 2: The four steps of our WSD-AF framework. Red arrows indicate optional steps in WSD-AF, while blue arrows represent traditional classification paths.

tic richness and reduce noise, and **Adaptive Fusion (AF)** to capture the dominant sentiment based on content. Details of our proposed approach are presented in Sec. 4.

4 WSD-AF Framework

Our WSD-AF framework is a plug-and-play method that integrates seamlessly with existing classification mechanisms and models to achieve more precise text classification. The overall workflow of the framework is illustrated in Fig.2 and detailed in Alg.1.

Firstly, we perform **dependency parsing**, a technique for analyzing grammatical relationships between words in a sentence. Through dependency parsing, we identify the subject, predicate, and object of a sentence to form a contextual sentence, which is then used for subsequent disambiguation.

The second step involves **word sense disambiguation (WSD)**. This process identifies the most contextually appropriate sense of a target word using WordNet. All potential senses of the target word are extracted from WordNet as candidates, and a pre-trained WSD model (e.g., MeMo-BERT-WSD) is used to predict the correct sense based on the target word's context. The predicted sense index is then mapped to WordNet's synsets to retrieve the semantic definition, providing a disambiguated interpretation of the target word. This approach combines knowledge-based techniques with deep learning, ensuring context-sensitive disambiguation.

In the third step, **adaptive fusion (AF)** is applied. We train an additional neural network, referred to as the fusion network, which takes two confidence vectors as input—one derived from the global sentence and the other from the context sentence provided to the classification model. The fusion network learns how global and local/contextual sentiments influence the model's output, effectively suppressing conflicting emotional noise and ensuring that the dominant sentiment is correctly identified.

In the final step, **final prediction**, the fusion network outputs a weighted vector, which is then processed using an `argmax` function to produce the final predicted label.

5 Experiment Setup

5.1 Models

In our experiments, we utilized BERT (Devlin, 2018) and RoBERTa (Liu, 2019) as the primary text classification models. For the WSD process, we employed the BERT-WSD model (Face, 2024) to extract fine-grained semantic representations, which were then integrated into our WSD-AF framework. The fusion model, is a lightweight neural network designed to dynamically weight and fuse the global (global vector) and contextual (context vector) sentiment representations. At its core, the fusion model incorporates a weight generation network (WeightNet) that determines the optimal weighting for combining the semantic vectors.

5.2 Datasets

We evaluated our approach using two distinct datasets: the **Sentiment Analysis Dataset (SAD)** and the **Hospital Reviews Dataset (HRD)**. The SAD consists of tweets with sentiment labels automatically generated based on the presence of sentiment. This dataset includes a total of $27,481$ samples for training and $3,534$ for testing. The HRD, on the other hand, comprises 951 sentiment-labeled reviews collected from Google Maps for hospitals in Bengaluru, India. Additionally, we **M**anually **C**reated a dataset **MC dataset** which consists of 150 noisy training samples and 60 testing samples. These samples were derived by prompting ChatGPT, ensuring that the format and style of the samples are consistent. The dataset is augmented to the existing dataset for a fair comparison. These datasets present a variety of challenges, including

noisy inputs and domain-specific expressions, providing a comprehensive evaluation of our approach.

5.3 Metrics

To evaluate the performance of our proposed framework, we utilized the following metrics:

Accuracy (Acc): This metric measures the proportion of correctly predicted instances out of the total number of instances. It provides a straightforward measure of overall model performance.

F1-Score (F1): The F1-score is the harmonic mean of precision and recall, providing a single measure that balances both metrics. It is particularly useful when dealing with imbalanced datasets.

These metrics collectively provide a comprehensive evaluation of the framework's performance across different aspects of classification quality.

5.4 Parameter Setting

The validation split ratio was set to 0.2 across all experiments. The general parameter settings included a total of 10 epochs, a batch size of 32, and a maximum sequence length of 128.

For the **SAD dataset**, during fine-tuning, $20,000$ samples were used for training, while $1,000$ samples were reserved for testing. For training the fusion model, $2,150$ samples from the SAD dataset and MC dataset were used for training, with 460 samples reserved for testing. The fusion model was trained with a batch size of 8 for 200 epochs. Finally, evaluation was performed on a separate set of 120 samples to assess model performance.

For the **HRD dataset**, during fine-tuning, 636 samples were used for training, while 200 samples were reserved for testing. For training the fusion model, 786 samples from the HRD dataset and MC dataset were used for training, with 460 samples reserved for testing. Similar to the SAD dataset, the fusion model was trained with a batch size of 8 for 200 epochs. Evaluation was conducted on a separate set of 260 samples to assess model performance.

The fine-tuned model weights and datasets used in this study are available at the following link: Fine-tuned Models and Datasets Repository.

6 Experiment Results and Discussion

Effect of fine-tuning. Fine-tuning pre-trained models like BERT effectively adapts general knowledge to specific tasks, making models more task-oriented. On the SAD dataset, fine-tuning improved accuracy from 35.87 to 85.43 by aligning pre-trained weights with task-specific patterns, enabling better handling of subtle sentiment cues in noisy or adversarial text. Without fine-tuning, the model often fails to capture task-relevant features.

Table 1: Comparison of classification performance of our methods (WSD, AF, and WSD-AF) with different models on the Sentiment Analysis Dataset. NFT denotes models that are not fine-tuned, while FT denotes fine-tuned models.

	NFT		FT	
Method	Acc↑	F1↑	Acc↑	F1↑
BERT	27.69	21.84	75.00	74.31
+ WSD	27.31	22.61	78.08	75.98
+ AF	27.53	21.61	77.42	75.29
+ WSD-AF	27.19	20.88	76.19	74.67
RoBERTa	62.31	47.84	79.02	77.09
+ WSD	62.31	47.84	77.31	75.31
+ AF	62.31	47.84	78.85	76.33
+ WSD-AF	62.31	47.84	77.02	75.27

Table 2: Comparison of classification performance of our methods on the Hospital Reviews Dataset.

	NFT		FT	
Method	Acc↑	F1↑	Acc↑	F1↑
BERT	35.00	26.40	83.33	83.47
+ WSD	34.17	25.51	80.00	79.85
+ AF	33.13	23.41	76.52	76.09
+ WSD-AF	32.17	15.66	75.00	74.31
RoBERTa	29.17	13.17	76.67	77.12
+ WSD	29.17	13.17	70.83	71.30
+ AF	29.17	13.17	75.83	76.07
+ WSD-AF	29.17	13.17	70.83	71.60

Comparison between WSD and baseline. Our method WSD, AF and WSD-AF achieves comparable results compared with baseline model, but still underperform the baseline. Trivially, disambiguating words provides richer semantics, which can reduce noise and improves accuracy. However, we attributes the low performance to the capabilities of target word recognition algorithm, where we disambiguate verbs, nouns so on, which is brute-force. Disambiguating these non-disambiguous words introduce unexpected noise and sentiment strength to the sentence, which interferes the prediction.

Comparison between AF and baseline. According to the experiment results, experiments with AF also underperforms, but better than WSD. We attribute the performance to the non-representative datasets. Training fusion model on partially noisy dataset can cause underperformance on a test dataset which is most noise-free.

Algorithm 1: WSD-AF Algorithm

Input: Training dataset $\mathcal{D}_{train}$, test dataset $\mathcal{D}_{test}$, pretrained model $\mathcal{M}_{pretrained}$
Output: Predicted labels for $\mathcal{D}_{test}$
Step 1: Fine-Tune Model
$\mathcal{M}_{fine} \leftarrow$ Fine-tune $\mathcal{M}_{pretrained}$ on $\mathcal{D}_{train}$;

Step 2: Train Fusion Model
Extract vectors from $\mathcal{D}_{train}$;
Train fusion model $\mathcal{F}$ using extracted vectors and labels;
Step 3: Target Word Recognition
for *each sentence* $s \in \mathcal{D}_{test}$ **do**
  Extract dominant phrases $\mathcal{P}$ using dependency parsing;
Step 4: Word Sense Disambiguation
for *each sentence* s *and phrases* $\mathcal{P} \in \mathcal{D}_{test}$ **do**
  Perform WSD on s using $\mathcal{P}$ to get disambiguated sentence s';
Step 5: Adaptive Fusion
for *each sentence* s *and* s' *in* $\mathcal{D}_{test}$ **do**
  Compute global vector $\mathbf{v}_g$ and context vector $\mathbf{v}_c$;
  Fuse vectors $\mathbf{v}_f \leftarrow \mathcal{F}(\mathbf{v}_g, \mathbf{v}_c)$;
Step 6: Final Prediction
Predict labels $\hat{y}$ using fused vectors $\mathbf{v}_f$;
Evaluate performance with $\mathcal{D}_{test}$ labels;

Comparison between WSD-AF and baseline. From the result, the performance is also a little worse. This is constrained by the limitation of AF and WSD. Combining the WSD and AF, we are convincd that WSD-AF has the potential to perform well on noisy dataset. Even we created a noisy dataset for our experiment, the quality and diversity is limited and cannot provide rich information for fusion model.

7 Limitations

This study has several limitations that warrant discussion. First, the approach requires significant computational resources, particularly when using large models like BERT large, which often exceed hardware capabilities and lead to GPU crashes. This constraint limited our experiments to smaller models like BERT base and RoBERTa base.

The most critical limitation, however, lies in the datasets used for evaluation. A truly representative dataset is essential for a comprehensive and fair evaluation. Ideally, such a dataset would include both clean sentences with clear sentiment information and sufficient noisy or adversarial inputs. Unfortunately, most available datasets predominantly contain straightforward text with explicit semantics, making it challenging to fully demonstrate the effectiveness of our framework. To address this gap, we used ChatGPT to generate the MC dataset; however, the generated data lacked diversity and volume. Furthermore, ChatGPT struggled to create realistic self-conflicting inputs, often producing fixed patterns that do not fully capture the complexity of real-world noisy data. In future work, the development of a more representative dataset with diverse and realistic noisy inputs would allow for a more robust and fair evaluation of our framework.

8 Conclusion

This study investigated the challenge of text classification in noisy and adversarial scenarios, focusing on implicit negation, self-conflicting semantics, and complex linguistic structures. Motivated by real-world applications, such as understanding nuanced emotional expressions, we proposed the WSD-AF framework, which integrates Word Sense Disambiguation for semantic enrichment and Adaptive Fusion for dynamic sentiment weighting.

WSD-AF achieved results worse than baseline models, because its performance was constrained by the lack of a representative dataset containing both clean and noisy data. Our MC dataset, generated using ChatGPT, lacked diversity and realistic adversarial examples, limiting the fusion model's potential. Additionally, WSD and AF introduced computational overhead, restricting scalability to larger models.

Future work should prioritize creating diverse, representative datasets and improving computational efficiency. Despite its limitations, WSD-AF highlights the importance of addressing noise and ambiguity in text classification and offers a promising foundation for further research.

9 Ethics Statement

This project adheres to ethical research practices. The datasets used, including SAD, HRD, and the manually crafted MC dataset, comply with licensing terms, and no personally identifiable information (PII) was included. The MC dataset was created solely for academic purposes to simulate noisy and adversarial inputs.

Our WSD-AF framework aims to improve text classification in noisy contexts, with applications in understanding nuanced expressions. We acknowledge potential risks like bias or misclassification and emphasize the need for responsible use within ethical and practical boundaries.

10 Statement of contributions

Mandy: Co-authored the abstract and introduction, developed the initial pipeline sketch finalized by Fengfei, and presented experimental results, including model performance and limitations. She designed the experiment setup, specifying models, datasets, metrics, and fine-tuning parameters. Mandy implemented the model fine-tuning code and handled dataset cleaning and loading to ensure efficient training.
Fengfei: Co-authored the abstract, motivation, and WSD-AF framework, structured the Overleaf file, and created all figures. He addressed limitations with noisy and conflicting inputs, proposing techniques for implicit negation and self-conflicting semantics. Fengfei developed and fine-tuned the fusion model, implemented all disambiguation-related code, and designed the main code structure, including dependency parsing and WordNet-based word sense disambiguation.
Danlin: Authored the related work section and collaborated on the introduction. She contributed to model training and fine-tuning on the HRD dataset. Danlin was also responsible for paper formatting and implemented the target word recognition code.

References

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hugging Face. 2024. Mime-memo/memo-bert-wsd. https://huggingface.co/MiMe-MeMo/MeMo-BERT-WSD. Accessed: 2024-12-18.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Do-Myoung Lee, Yeachan Kim, and Chang gyun Seo. 2022. Context-based virtual adversarial training for text classification with noisy labels.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Hardik Meisheri, Kunal Ranjan, and Lipika Dey. 2017. Sentiment extraction from consumer-generated noisy short texts. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 399–406.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2021. Adversarial training methods for semi-supervised text classification.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream nlp applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey.