



Fine-Tuning BERT for Context-Aware Phrasal Verb Disambiguation

Sujair MZIAA. Ibrahim^a, Abishethvarman Vadivel^{b*}, Kumara TGS. Banage^c

^{a,c} Department of Software Engineering, Sabaragamuwa University of Sri Lanka, 70140

^b Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, 70140

^aEmail: ssujair@gmail.com

^bEmail: abishethvarman@gmail.com

^cEmail: kumara@appsc.sab.ac.lk

Abstract

Transformer-based models like BERT have significantly advanced natural language processing by enabling deep contextualized language representations. While powerful in token prediction and structural understanding, their success does not inherently imply a strong grasp of semantic meaning, particularly for ambiguous expressions. Semantic disambiguation is a core challenge, especially with phrasal verbs, whose meanings are often non-compositional and highly context-dependent, making them difficult for models trained primarily on surface-level patterns. This research addresses the gap where foundational BERT models are not optimally designed for meaning classification. We investigate whether fine-tuning BERT can improve its performance in phrasal verb meaning classification. To this end, we created a context-rich dataset focused on phrasal verb semantic interpretation, augmenting it with LLM-generated sentences for robustness. We compare a baseline bert-base-uncased model with a fine-tuned version, utilizing QLoRA for efficient training given hardware constraints. Model performance before and after fine-tuning is evaluated using a diverse set of metrics including cosine similarity, BLEU, ROUGE-L, Jaccard similarity, and METEOR. The results consistently demonstrate that fine-tuning significantly enhances the model's ability to capture the semantic intent behind phrasal verbs, confirming the effectiveness of contextual specialization through targeted training for nuanced language understanding. The code implementation can be found at: <https://github.com/Abishethvarman/bert-semantic-classification>

Keywords: BERT Fine-tuning; Natural Language Processing; Phrasal Verbs; QLoRA; Semantic Disambiguation; Transformer Models.

* Corresponding author. Email address: abishethvarman@gmail.com

1. Introduction

1.1. Introduction

Transformer-based models such as BERT have significantly advanced natural language processing by enabling deep contextualized representations of language. These models use self-attention mechanisms to capture relationships between words, allowing them to predict masked tokens with high accuracy across various linguistic tasks. However, success in token prediction does not necessarily imply a strong grasp of meaning, particularly when the target is to understand semantic relationships rather than just surface-level patterns.

A core principle underlying large language models (LLMs) is the task of next-token prediction: given a sequence of tokens, the model learns to predict the most likely next token using statistical patterns learned during pre-training [1], [2], [3]. This deceptively simple objective enables models to acquire deep semantic representations of language. The Transformer architecture [4], through its self-attention mechanism, allows LLMs to capture contextual dependencies across long text spans, facilitating nuanced understanding of meaning, syntax, and discourse. Models such as BERT [5] and T5 [6] demonstrate how token-level objectives like masked language modeling or sequence-to-sequence learning can yield general-purpose semantic representations transferable across tasks.

More recent models—including PaLM [7], LLaMA 2 [8], GPT-4 [9], Gemini [10], [11], and Gemma Team [12]—extend these principles to larger scales and modalities, learning rich, context-sensitive representations that support in-context learning, reasoning, and even multimodal understanding. Token prediction, though local in nature, has proven remarkably effective for capturing global semantics, enabling LLMs to generalize across languages, tasks, and domains.

Semantic disambiguation is a central challenge in language understanding. It involves identifying the correct meaning of a word or phrase depending on its surrounding context. This task becomes especially difficult in the

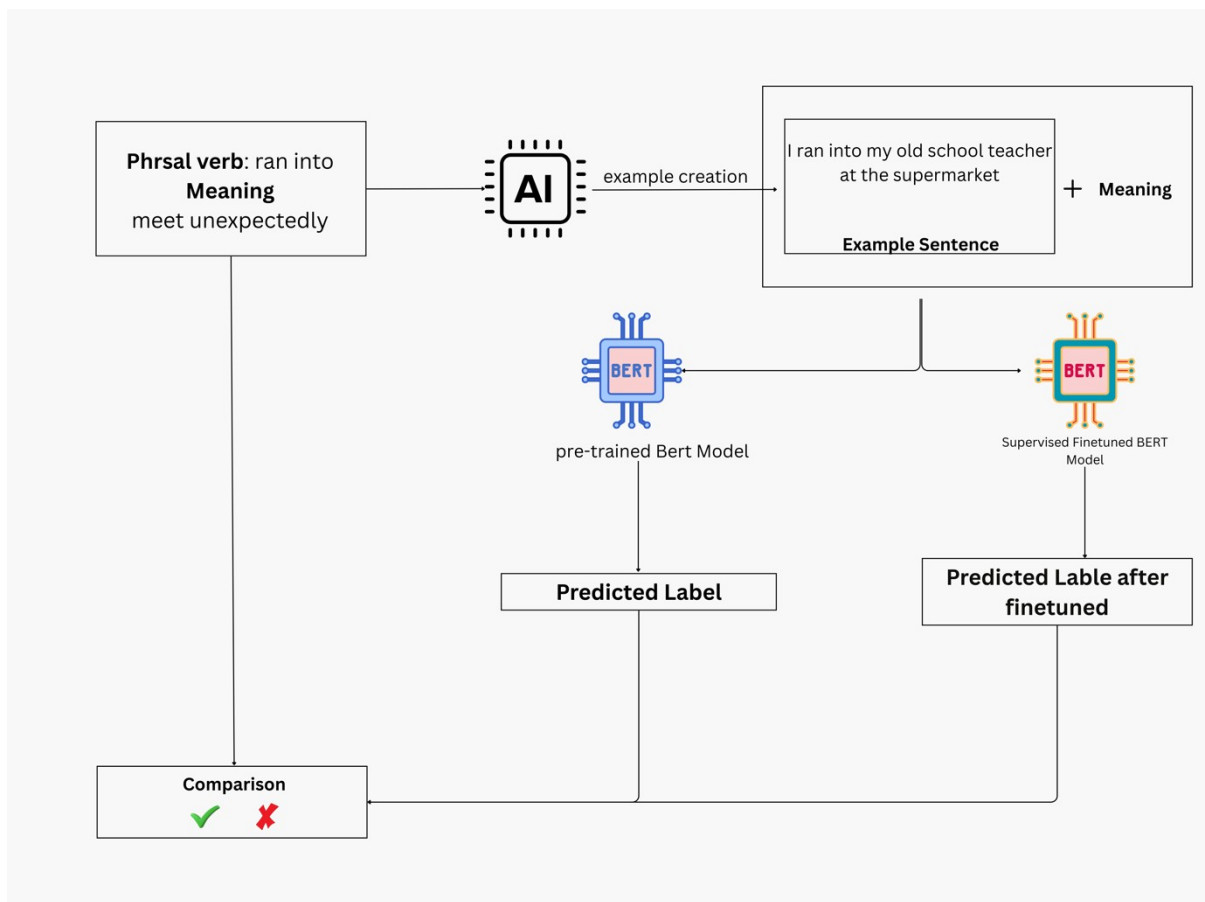


Figure 1: Logical Flow

case of phrasal verbs, which often carry meanings that are not directly derived from their individual components. For example, the expression “run out” can imply exhaustion of supplies or someone physically leaving a place, depending on the sentence. This context-dependence makes them hard to interpret using basic lexical features.

Phrasal verbs are known for their high semantic ambiguity. Their meanings vary widely based on usage, and such variability is not usually captured well by models trained primarily for token-level prediction. BERT is designed to generate context-aware representations, but it is not explicitly trained to classify the intended meaning of ambiguous expressions. This highlights a key gap: the foundational BERT model, although powerful in encoding syntax and structure, does not optimally distinguish between different meanings of the same phrase in different contexts.

To address this gap, we investigate whether supervised fine-tuning (SFT) with parameter-efficient QLoRA adapters can enhance BERT’s ability to perform context-aware semantic classification of phrasal verbs. The task is framed as a two-phase process. In the first phase (Bare flow), a pre-trained BERT model is used without task-specific fine-tuning to detect phrasal verbs in sentences and assign meaning labels. In the second phase (SFT Flow), BERT is augmented with QLoRA adapters and fine-tuned on a curated, multi-sense phrasal verb dataset, enabling it to learn nuanced distinctions between literal and idiomatic usages. This design allows a direct comparison between the zero-shot baseline and the fine-tuned model, quantifying the impact of targeted training on semantic disambiguation. Model performance is evaluated before and after fine-tuning using multiple metrics, including cosine similarity [13], BLEU [14], ROUGE-L [15], Jaccard similarity [16], and METEOR [17].

This research addresses two questions: how effectively can a pre-trained BERT model classify phrasal verbs in context, and to what extent does supervised fine-tuning improve [18] its detection and classification performance? By focusing on phrasal verbs as a representative case of semantic ambiguity, this work evaluates BERT’s potential for broader context-aware classification tasks and lays the groundwork for extending the approach to idioms and other multi-word expressions.

1.2. Relevant Works

Phrasal verbs, such as run into, break down, and take off, are highly ambiguous multi-word expressions whose meanings can shift dramatically depending on context. Their interpretation often requires not only syntactic parsing but also an understanding of idiomatic usage and subtle semantic cues, making them difficult for NLP systems to process accurately. Similar challenges are well-documented in other linguistic constructions, including complex nominals (CNs), genitives, and adjective–noun phrases, which display non-compositionality, context dependence, and multiple possible semantic relations [19], [20], [21]. For example, the meaning of “musical clock” differs greatly from “musical creation,” and genitive constructions like “John’s car” can signal ownership or authorship depending on context. Such variability is mirrored in phrasal verbs, where a single expression like run out may mean to deplete or to exit, depending on its usage.

Semantic disambiguation — the task of determining the intended meaning of a word or phrase based on context is central to effective language understanding. Early research on semantic relation classification often relied on hand-crafted rules [22], [23] or statistical methods exploiting lexico syntactic patterns [24], [25]. While these approaches offered initial insights, they struggled to scale to the variability and ambiguity present in natural language. More recent work has explored supervised and semi-supervised machine learning for disambiguating semantic relations in noun phrases and other constructions [21], [26], [27], [28], [29], but accuracy is often limited by errors in automatic sense disambiguation and by missing or incomplete feature combinations in training data [19].

The introduction of transformer-based models, particularly BERT [30], marked a significant step forward in NLP. Through its masked language modeling and next-sentence prediction objectives, BERT captures deep bidirectional contextual representations, enabling strong performance on a variety of tasks. However, as several studies have noted [19], [28] strong token prediction and syntactic modeling do not necessarily imply robust semantic disambiguation, particularly for multi-word expressions. Foundational BERT[30] models are not explicitly trained to identify and classify the specific sense of an ambiguous expression in varied contexts, which can limit their performance for tasks like phrasal verb classification without additional task-specific fine-tuning.

While research on phrasal verbs in NLP is relatively limited compared to other linguistic phenomena, existing studies have examined approaches such as statistical co-occurrence analysis, translation-based sense alignment, and lexicon-enhanced neural models [21], [28]. These methods highlight the importance of combining contextual embeddings with sense inventories or semantic lexicons. Yet, the integration of parameter-efficient fine-tuning techniques with transformer models for phrasal verb meaning classification remains underexplored.

This study addresses this gap by evaluating BERT’s ability to classify phrasal verbs with correct meanings in context and by investigating whether supervised fine-tuning with parameter-efficient QLoRA[31] adapters can enhance its performance on this task. By using phrasal verbs as a representative case of semantic ambiguity, our work extends prior research on semantic relation classification and explores targeted fine-tuning [32], [33], [34], [35], [36] as a means of improving contextual semantic understanding.

2. Materials and Methods

2.1. Dataset Preparation

The dataset used in this study was sourced from the openly accessible repository “generated-english-phrasal-verbs” on GitHub, which compiles structured semantic data on over 3,350 English phrasal verbs. Each phrasal verb entry is enriched with linguistic details including morphological derivatives (e.g., “abides by”, “abided by”), a concise description representing its semantic meaning, and synonym sets. The presence of both grammatical variants and contextual illustrations makes this resource particularly well-suited for semantic disambiguation tasks.

For the purpose of our study, we focused specifically on two components: the base phrasal verb and its description as a proxy for semantic meaning. with the meaning we have created examples for each phrasal verbs using LLM and This resulted in a curated dataset containing 11,894 labeled instances, each comprising a contextual sentence where a specific phrasal verb occurs and the corresponding human-written semantic description. These labels serve as ground truth meanings, enabling the model to learn not merely surface-level token patterns, but the deeper intent embedded in real-world usage.

Unlike conventional classification datasets with fixed categories, this dataset exhibits rich semantic diversity, with many phrasal verbs sharing similar surface forms but diverging in meaning based on context. This characteristic directly supports the core objective of our research: to evaluate how well a transformer-based language model like BERT can be fine-tuned to capture the contextual semantics of phrasal verbs. By anchoring each sentence to a meaning label, the dataset transforms the inherently ambiguous problem of phrasal verb understanding into a supervised semantic classification task grounded in context.

2.2. Methodology

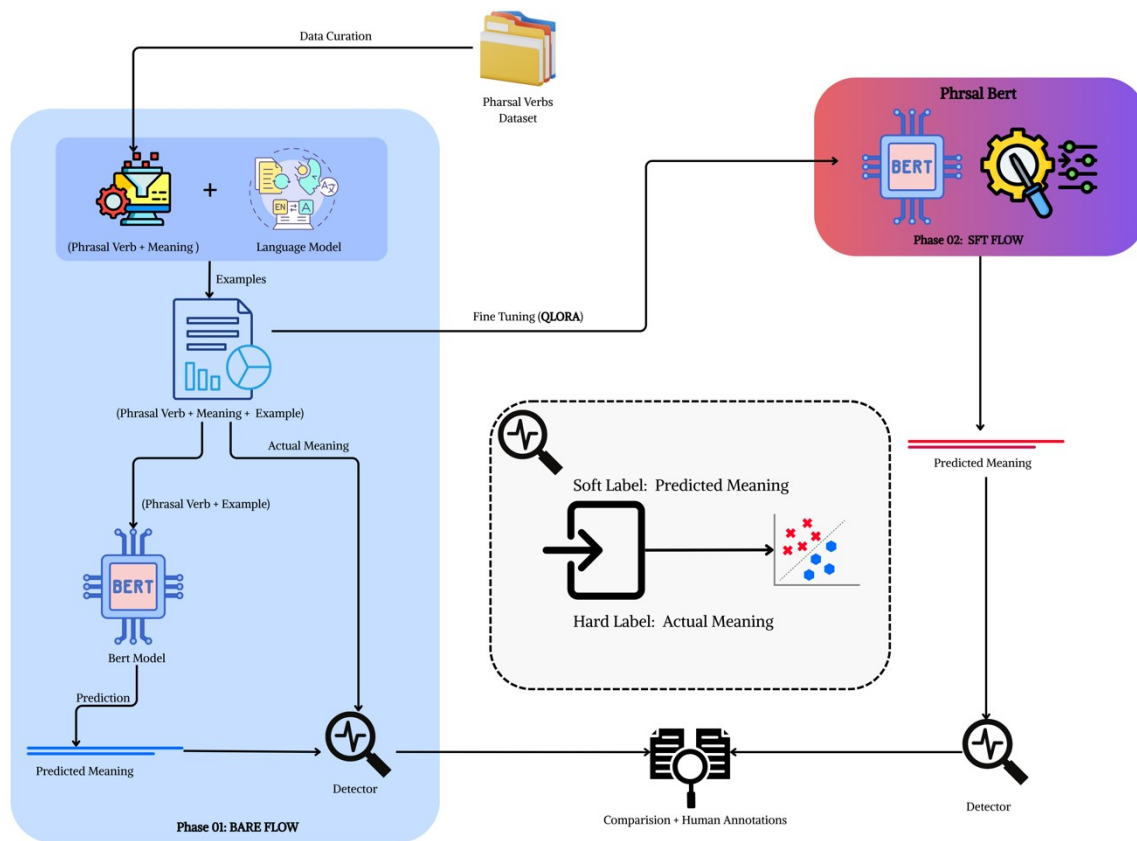


Figure 2: Methodology

This section details the methodology employed to investigate the semantic understanding of phrasal verbs using BERT and its fine-tuned variants. The primary objective was to evaluate whether fine-tuning the base BERT model could enhance its ability to discern the contextual meanings of phrasal verbs. The methodology encompasses dataset construction, preprocessing, model architecture and training, evaluation strategies, and infrastructure considerations, with a focus on efficiency using QLoRA.

The core dataset used in this study was obtained from an open-source GitHub generated-english-phrasal-verbs repository. The file `phrasal.verbs.build.jsonl` contained structured data on 3,350 phrasal verbs, including their variants, semantic descriptions, synonyms, and translations. For this work, only the essential fields relevant to our objective were extracted: the phrasal verb and its semantic description (used as the gold-standard meaning). Contextual examples for each phrasal verb were generated using large language models (LLMs), with prompts explicitly including both the phrasal verb and its reference meaning to ensure semantic fidelity. This extraction resulted in a dataset of 11,894 instances of phrasal verbs, each comprising a sentence, its associated phrasal verb, and a labeled meaning. The rich contextual and semantic diversity of the dataset made it particularly suitable for both zero-shot evaluation and supervised fine-tuning.

Each instance in the dataset was formatted as an input–output pair, where the input was a sentence containing a phrasal verb, and the output was its corresponding semantic meaning. To improve contextual diversity and robustness, additional sentences were generated using a large language model. These augmentations enabled the evaluation of model performance not only on observed linguistic patterns but also on varied or unseen constructions.

Two experimental flows were defined: (1) Bare Flow: The pre-trained bert-base-uncased model from the Hugging Face Transformers library was evaluated without task-specific fine-tuning. A classification head was trained on top of the frozen BERT embeddings to assign semantic labels based on sentence context. This setup assessed BERT’s zero-shot capability for phrasal verb meaning classification. (2) SFT Flow: The same base model was fine-tuned using QLoRA adapters, allowing the model to learn distinctions between literal and idiomatic usages while retaining most pre-trained parameters.

To evaluate semantic alignment between predicted and gold-standard meanings, several complementary evaluation metrics were employed: cosine similarity (using all-MiniLM-L6-v2 embeddings), BLEU, METEOR, ROUGE-L, and Jaccard similarity. Library versions were: NLTK v3.9.1 for BLEU/METEOR, rouge v1.0.1 for ROUGE-L, scikit-learn v1.7.0 for Jaccard and cosine similarity, and sentence-transformers v4.1.0 for embedding-based similarity. Data handling used pandas v2.3.0 and numpy v2.3.1, while PyTorch v2.7.1 was used for model computations. Evaluation followed a dual approach: (i) exact-match accuracy for strict correctness, and (ii) averaged similarity metrics for partial correctness.

To address the limitations observed in the Bareflow results, the study employed QLoRA (Quantized Low-Rank Adapter) for efficient fine-tuning. QLoRA enhances transformer models by introducing low-rank, trainable adapter layers while keeping the original model weights frozen and quantized to 4-bit precision. This approach significantly reduces memory usage, allowing large models to be fine-tuned even on hardware with limited computational resources. The training was conducted with a learning rate of $2e-4$, selected after empirical tuning. Due to memory constraints, a small batch size was used in conjunction with gradient accumulation to simulate larger effective batches. A cosine learning rate scheduler with a warm-up phase was applied to stabilize convergence. The training ran for three epochs, and only the QLoRA adapter layers and classifier head were updated during training, preserving the general language understanding of the base BERT model while tailoring it to the task of phrasal verb semantic classification.

Following fine-tuning (SFT Flow), the evaluation pipeline was re-applied to the updated model. The results demonstrated clear improvements across all semantic similarity metrics. BLEU scores showed particular gains for literal phrasal verbs, while METEOR and cosine similarity scores improved notably for idiomatic and context-dependent usages. ROUGE-L and Jaccard similarity also reflected enhanced structural and vocabulary-level alignment. These findings confirmed that the QLoRA-fine-tuned BERT model had developed a more nuanced understanding of sentence context and could more accurately predict intended meanings, especially when a phrasal verb had multiple interpretations based on its usage.

All training and evaluation were conducted on an Apple Silicon M3 Pro machine lacking native CUDA support.

Despite these limitations, QLoRA enabled efficient training by reducing memory demands. Although the bitsandbytes library was not fully compatible with the Mac architecture and the PyTorch MPS backend had some limitations, training was successful with reduced batch sizes and extended gradient accumulation. Logs and TensorBoard visualizations confirmed stable convergence patterns for both Bare Flow and SFT Flow experiments. The final model demonstrated strong generalization to unseen examples, indicating that task-specific fine-tuning using QLoRA significantly enhances BERT’s capability to semantically disambiguate phrasal verbs in diverse linguistic contexts.

3. Results

The performance of the base BERT model in a zero-shot setting and the fine-tuned BERT model was systematically evaluated to understand their capabilities in semantic disambiguation of phrasal verbs. The zero-shot BERT model, which had not undergone any task-specific fine-tuning, served as a baseline to measure the intrinsic semantic understanding of phrasal verbs by the transformer architecture. Despite its powerful contextual encoding ability, the zero-shot model exhibited limited accuracy when predicting the precise meaning of phrasal verbs in varied contexts. This limitation is attributable to BERT’s original training objective, which is primarily focused on predicting masked tokens rather than directly learning semantic classifications. Consequently, its token-level predictions do not always translate effectively into fine-grained semantic distinctions, especially for polysemous expressions such as phrasal verbs whose meanings vary drastically with context.

Table 1: Performance Metrics Before and After Fine-Tuning BERT

Benchmark Metrics	Before Fine-Tuning	After Fine-Tuning
Cosine Similarity	0.5889	0.6189
BLUE Score	0.2570	0.3150
ROUGE-L	0.4623	0.4901
Jaccard Similarity	0.3484	0.3648
METEOR	0.3555	0.3607

Fine-tuning BERT with our curated and augmented phrasal verb dataset using QLoRA substantially improved the model's ability to assign correct semantic meanings. The adaptation of low-rank adapter layers allowed the model to specialize in this task without the need for full retraining of the enormous BERT weights, thus optimizing computational efficiency while enhancing semantic understanding. Post fine-tuning, the model showed notable increases in performance metrics such as BLEU, METEOR, cosine similarity, ROUGE-L, and Jaccard similarity, reflecting better alignment with the gold-standard semantic labels. For example, the cosine similarity improved from 0.5889 to 0.6189, BLEU score rose from 0.2570 to 0.3150, ROUGE-L increased from 0.4623 to 0.4901, Jaccard similarity moved up from 0.3484 to 0.3648, and METEOR enhanced from 0.3555 to 0.3607. These improvements were particularly evident when the model dealt with idiomatic or less literal usages of phrasal verbs, where contextual cues are paramount.

A detailed comparison of seen versus unseen phrasal verbs during evaluation highlighted the generalization capacity of the fine-tuned model. For phrasal verbs present in the training data (seen verbs), the model achieved higher accuracy, indicating effective memorization and contextual embedding learning. However, more importantly, the fine-tuned BERT demonstrated encouraging results on unseen phrasal verbs, suggesting that the adapter-based training imbued the model with a deeper semantic compositionality and the ability to leverage context effectively even for novel expressions. This ability to generalize underscores the success of using context-rich examples and a focused fine-tuning approach.

Further analysis comparing phrasal verbs with similar meanings versus those with dissimilar meanings revealed subtle nuances in the model's semantic discrimination power. The fine-tuned BERT was generally adept at differentiating between closely related senses, which is critical given the often overlapping semantic fields of phrasal verbs. Nevertheless, some confusions persisted in cases where the semantic boundaries were very fine or when contextual information was minimal or ambiguous. These cases provide valuable insights into the limitations of current transformer architectures for nuanced semantic tasks and indicate areas for future research, such as integrating external knowledge bases or improved contextual embeddings.

Qualitative examination of model predictions illuminated typical success and failure cases. Correct predictions often occurred when the sentence context provided clear disambiguating cues, enabling the model to focus on relevant syntactic and semantic features. Conversely, errors tended to arise in sentences with figurative language, idiomatic expressions with rare usage, or ambiguous syntax. These qualitative insights reinforce the quantitative findings and demonstrate the critical role of fine-tuning in equipping BERT to handle the diverse and complex nature of phrasal verb semantics.

In summary, while BERT's foundational transformer architecture is inherently strong in language representation due to its masked token prediction pretraining, Overall, the results confirm that while pre-trained BERT can capture contextual meaning, task-specific fine-tuning is essential for achieving high accuracy in phrasal verb semantic classification. By combining BERT's context-aware representations with efficient adapter-based tuning via QLoRA, the model bridged the gap between general language understanding and precise semantic interpretation, achieving notable improvements with minimal computational overhead.

4. Conclusions

This study addressed the gap in BERT's ability to accurately capture the contextual meanings of phrasal verbs, a challenge arising from their idiomatic and non-compositional nature. We investigated whether supervised fine-tuning with QLoRA could improve semantic classification performance. The fine-tuned model demonstrated clear gains over the base BERT across multiple evaluation metrics, including BLEU, ROUGE-L, and cosine similarity. These results confirm that lightweight supervised adaptation can effectively enhance transformer models for context-aware interpretation of phrasal verbs, contributing to more accurate and semantically robust NLP systems.

5. Acknowledgments

5.1. Limitations

This study, while offering valuable insights into the Supervised fine-tuning of BERT for phrasal verb meaning classification, is subject to several limitations that warrant consideration for future research. Firstly, the training and evaluation were conducted on a relatively small dataset of phrasal verbs. While carefully constructed to be context-rich, a larger and more diverse dataset could further enhance the model's robustness and generalizability across a wider array of linguistic patterns and infrequent phrasal verb usages. Secondly, the research was exclusively conducted on the English language. Phrasal verbs and semantic ambiguities present unique challenges across different languages; therefore, the findings may not directly translate to other languages without further investigation, particularly those with distinct grammatical structures or idiomatic expressions. Lastly, the primary focus of this study was limited to the semantic disambiguation of phrasal verbs. While a critical area of linguistic ambiguity, the generalizability of these fine-tuning approaches to other complex semantic phenomena, such as idioms, metaphors, or highly polysemous single words, was not explored. Future work could extend this methodology to a broader spectrum of ambiguous linguistic expressions.

6. Reference

- [1] A. R. Openai, K. N. Openai, T. S. Openai, and I. S. Openai, "Improving language understanding by generative pre-training," 2018, Accessed: Aug. 20, 2025. [Online]. Available: <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *storage.prod.researchhub.com* A Radford, J Wu, R Child, D Luan, D Amodei, I Sutskever OpenAI blog, 2019•*storage.prod.researchhub.com*, Accessed: Aug. 20, 2025. [Online]. Available: <https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf>
- [3] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv Neural Inf Process Syst*, vol. 33, pp. 1877–1901, 2020, Accessed: Aug. 20, 2025. [Online]. Available: <https://commoncrawl.org/the-data/>
- [4] A. Vaswani *et al.*, "Attention is all you need," *proceedings.neurips.cc*, Accessed: Aug. 10, 2025. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

- [5] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *aclanthology.org* J Devlin, MW Chang, K Lee, K Toutanova *Proceedings of the 2019 conference of the North American chapter of, 2019•aclanthology.org*, pp. 4171–4186, Accessed: Aug. 20, 2025. [Online]. Available: https://aclanthology.org/N19-1423/?utm_campaign=The+Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-m9bbH_7ECE1h3IZ3D61TYg52rKpifVNjL4fvJ85uqggrXsWDBTB7YooFLJeNXHWqhVoyC
- [6] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *jmlr.org* C Raffel, N Shazeer, A Roberts, K Lee, S Narang, M Matena, Y Zhou, W Li, PJ Liu *Journal of machine learning research, 2020•jmlr.org*, vol. 21, pp. 1–67, 2020, Accessed: Aug. 20, 2025. [Online]. Available: <http://www.jmlr.org/papers/v21/20-074.html>
- [7] A. Chowdhery *et al.*, “Palm: Scaling language modeling with pathways,” *jmlr.org* A Chowdhery, S Narang, J Devlin, M Bosma, G Mishra, A Roberts, P Barham, HW Chung *Journal of Machine Learning Research, 2023•jmlr.org*, vol. 24, pp. 1–113, 2023, Accessed: Aug. 20, 2025. [Online]. Available: <http://www.jmlr.org/papers/v24/22-1144.html>
- [8] H. Touvron, L. Martin, K. Stone, ... P. A. preprint arXiv, and undefined 2023, “Llama 2: Open foundation and fine-tuned chat models,” *arxiv.org* H Touvron, L Martin, K Stone, P Albert, A Almahairi, Y Babaei, N Bashlykov, S Batra *arXiv preprint arXiv:2307.09288, 2023•arxiv.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2307.09288>
- [9] J. Achiam, S. Adler, S. Agarwal, ... L. A. preprint arXiv, and undefined 2023, “Gpt-4 technical report,” *arxiv.org* J Achiam, S Adler, S Agarwal, L Ahmad, I Akkaya, FL Aleman, D Almeida, J Altenschmidt *arXiv preprint arXiv:2303.08774, 2023•arxiv.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2303.08774>
- [10] G. Team, R. Anil, S. Borgeaud, J. Alayrac, ... J. Y. preprint arXiv, and undefined 2023, “Gemini: a family of highly capable multimodal models,” *arxiv.org* G Team, R Anil, S Borgeaud, JB Alayrac, J Yu, R Soricut, J Schalkwyk, AM Dai, A Hauthar *arXiv preprint arXiv:2312.11805, 2023•arxiv.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2312.11805>
- [11] G. Team, P. Georgiev, V. Lei, R. Burnell, ... L. B. preprint arXiv, and undefined 2024, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arxiv.org* G Team, P Georgiev, VI Lei, R Burnell, L Bai, A Gulati, G Tanzer, D Vincent, Z Pan, S Wang *arXiv preprint arXiv:2403.05530, 2024•arxiv.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2403.05530>
- [12] G. Team, T. Mesnard, C. Hardin, ... R. D. preprint arXiv, and undefined 2024, “Gemma: Open models based on gemini research and technology,” *arxiv.org* G Team, T Mesnard, C Hardin, R Dadashi, S Bhupatiraju, S Pathak, L Sifre, M Rivière *arXiv preprint arXiv:2403.08295, 2024•arxiv.org*, 2024, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2403.08295>
- [13] H. Steck, C. Ekanadham, and N. Kallus, “Is cosine-similarity of embeddings really about similarity?,” *dl.acm.org*, pp. 887–890, May 2024, doi: 10.1145/3589335.3651526.
- [14] M. Post, “A Call for Clarity in Reporting BLEU Scores,” *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference*, vol. 1, pp. 186–191, 2018, doi: 10.18653/v1/w18-6319.
- [15] C. Lin, F. O.-N. workshop, and undefined 2004, “Looking for a few good metrics: ROUGE and its evaluation,” *research.nii.ac.jp*, 2004, Accessed: Aug. 20, 2025. [Online]. Available: https://research.nii.ac.jp/ntcir/ntcir-ws4/NTCIR4-WN/OPEN/OPENSUB_Chin-Yew_Lin.pdf

- [16] S. Bag, S. Kumar, M. T.-I. Sciences, and undefined 2019, "An efficient recommendation generation using relevant Jaccard similarity," *Elsevier*, vol. 483, pp. 53–64, 2019, doi: 10.1016/j.ins.2019.01.023.
- [17] S. Banerjee, A. L.-P. of the acl workshop on intrinsic, and undefined 2005, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," *aclanthology.org*, pp. 376–380, Accessed: Aug. 20, 2025. [Online]. Available: <https://aclanthology.org/W05-0909.pdf>
- [18] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "SUPERVISED CONTRASTIVE LEARNING FOR PRE-TRAINED LANGUAGE MODEL FINE-TUNING," *ICLR 2021 - 9th International Conference on Learning Representations*, 2021, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/pdf/2011.01403>
- [19] D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju, "Models for the semantic classification of noun phrases," *aclanthology.org* D Moldovan, A Badulescu, M Tatu, D Antohe, R Girju *Proceedings of the Computational Lexical Semantics Workshop at HLT, 2004*•*aclanthology.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://aclanthology.org/W04-2609.pdf>
- [20] M. L.-C. Linguistics and undefined 2002, "The disambiguation of nominalizations," *direct.mit.edu* M Lapata *Computational Linguistics*, 2002•*direct.mit.edu*, Accessed: Aug. 10, 2025. [Online]. Available: <https://direct.mit.edu/coli/article-abstract/28/3/357/1761>
- [21] B. Rosario, M. H.-P. of the 2001 C. on, and undefined 2001, "Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy," *aclanthology.org* B Rosario, MA Hearst *Proceedings of the 2001 Conference on Empirical Methods in Natural*, 2001•*aclanthology.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://aclanthology.org/W01-0511.pdf>
- [22] T. Finin, *The semantic interpretation of compound nominals*. 1980. Accessed: Aug. 10, 2025. [Online]. Available: <https://search.proquest.com/openview/b78539371ace99efb317ff65081ea0f6/1?pq-origsite=gscholar&cbl=18750&diss=y>
- [23] J. N. . Levi, "The syntax and semantics of complex nominals," p. 301, 1978.
- [24] M. Lauer, M. D. preprint cmp-lg/9409003, and undefined 1994, "A probabilistic model of compound nouns," *arxiv.org* M Lauer, M Drasar *Xiv preprint cmp-lg/9409003*, 1994•*arxiv.org*, Accessed: Aug. 10, 2025. [Online]. Available: <https://arxiv.org/abs/cmp-lg/9409003>
- [25] L. V.-C. 1994 V. 2: T. 15th International and undefined 1994, "Algorithm for automatic interpretation of noun sequences," *aclanthology.org* L Vanderwende *COLING 1994 Volume 2: The 15th International Conference on*, 1994•*aclanthology.org*, Accessed: Aug. 10, 2025. [Online]. Available: <https://aclanthology.org/C94-2125.pdf>
- [26] R. Girju, A. Badulescu, D. M.-C. Linguistics, and undefined 2006, "Automatic discovery of part-whole relations," *direct.mit.edu* R Girju, A Badulescu, D Moldovan *Computational Linguistics*, 2006•*direct.mit.edu*, 2006, Accessed: Aug. 20, 2025. [Online]. Available: <https://direct.mit.edu/coli/article-abstract/32/1/83/1901>
- [27] R. Girju, A.-M. Giuglea, M. Olteanu, O. Fortu, O. Bolohan, and D. Moldovan, "Support vector machines applied to the classification of semantic relations in nominalized noun phrases," *aclanthology.org* R Girju, AM Giuglea, M Olteanu, O Fortu, O Bolohan, D Moldovan *Proceedings of the Computational Lexical Semantics Workshop at HLT, 2004*•*aclanthology.org*, Accessed: Aug. 10, 2025. [Online]. Available: <https://aclanthology.org/W04-2610.pdf>
- [28] R. Navigli, P. Velardi, U. Di, R. La Sapienza, and A. Gangemi, "Ontology learning and its application to automated terminology translation," *ieeexplore.ieee.org* R Navigli, P Velardi, A Gangemi *IEEE Intelligent systems*, 2003•*ieeexplore.ieee.org*, 2003, Accessed: Aug. 10, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/1179190/>

- [29] B. Rosario, M. Hearst, ... C. F. the 40th A. M. of, and undefined 2002, "The descent of hierarchy, and selection in relational semantics," *aclanthology.org* B Rosario, MA Hearst, CJ Fillmore *Proceedings of the 40th Annual Meeting of the Association for, 2002*•*aclanthology.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://aclanthology.org/P02-1032.pdf>
- [30] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "Bert: Pre-training of deep bidirectional transformers for language understanding," *aclanthology.org* J Devlin, MW Chang, K Lee, K Toutanova *Proceedings of the 2019 conference of the North American chapter of, 2019*•*aclanthology.org*, pp. 4171–4186, Accessed: Jun. 09, 2025. [Online]. Available: https://aclanthology.org/N19-1423/?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-_m9bbH_7ECE1h3lZ3D61TYg52rKpifVNjL4fvJ85uqggrXsWDBTB7YooFLJeNXHWqhvOyC
- [31] T. Dettmers, A. Pagnoni, ... A. H.-A. in neural, and undefined 2023, "Qlora: Efficient finetuning of quantized llms," *proceedings.neurips.cc* T Dettmers, A Pagnoni, A Holtzman, L Zettlemoyer *Advances in neural information processing systems, 2023*•*proceedings.neurips.cc*, Accessed: Jun. 09, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html
- [32] K. Church, Z. Chen, Y. M.-N. L. Engineering, and undefined 2021, "Emerging trends: A gentle introduction to fine-tuning," *cambridge.org*, vol. 27, pp. 763–778, 2021, doi: 10.1017/S1351324921000322.
- [33] Z. Fu, H. Yang, A. So, W. Lam, ... L. B.-P. of the A., and undefined 2023, "On the effectiveness of parameter-efficient fine-tuning," *ojs.aaai.org*, 2023, Accessed: Aug. 20, 2025. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/26505>
- [34] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," *Springer*, vol. 11856 LNAI, pp. 194–206, 2019, doi: 10.1007/978-3-030-32381-3_16.
- [35] Y. G. Xu, X. P. Qiu, L. G. Zhou, and X. J. Huang, "Improving bert fine-tuning via self-ensemble and self-distillation," *Springer*, vol. 38, no. 4, pp. 853–866, Jul. 2023, doi: 10.1007/S11390-021-1119-0.
- [36] T. Zhang, F. Wu, A. Katiyar, ... K. W. preprint arXiv, and undefined 2020, "Revisiting few-sample BERT fine-tuning," *arxiv.org*, Accessed: Aug. 20, 2025. [Online]. Available: <https://arxiv.org/abs/2006.05987>