DeBERTa at SemEval-2024 Task 9: Using DeBERTa for Defying Common Sense

Marco Siino
Department of Electrical, Electronic
and Computer Engineering
University of Catania
Italy
marco.siino@unipa.it

Abstract

The significant achievements of language models have motivated researchers in the natural language processing (NLP) community to confront challenges requiring nuanced and implicit reasoning, inspired by human-like common-sense understanding. Although efforts focusing on vertical thinking tasks have received substantial recognition, there remains a notable lack of investigation into lateral thinking puzzles. To bridge this void, the authors at SemEval-2024 propose BRAINTEASER: a multiple-choice Question Answering task designed meticulously to assess the model's lateral thinking capabilities and its capacity to question default common-sense assumptions. Specifically, at the SemEval-2024 Task 9, for the first subtask (i.e., Sentence Puzzle) the organizers asked the participants to develop models able to reply to multi-answer brain-teasing questions. For this purpose, we propose the application of a DeBERTa model in a zero-shot configuration. The proposed approach achieves an aggregate score of 0.250. Suggesting a significant room for improvements in future works.

1 Introduction

Human reasoning encompasses two fundamental types of cognitive processing: vertical and lateral thinking. Vertical thinking is marked by its sequential and analytical approach, drawing upon principles of rationality, logic, and rule-following, often attributed to the left-brain hemisphere (Knauff, 2013; Huang et al., 2023). This mode of thinking is essential for creating logical pathways, such as understanding physical scenarios or solving riddles based on direct associations. In contrast, lateral thinking, often referred to as "thinking outside the box," is a creative cognitive process. It entails exploring problems from unconventional perspectives and challenging ingrained assumptions. Lateral thinking, associated with the right-brain hemisphere, is crucial for resolving unconventional puzzles by defying common-sense associations and considering alternative perspectives.

While natural language processing (NLP) models have made significant strides in vertical thinking tasks, particularly in the field of large language models (LLMs). Their performance in lateral thinking remains largely unexplored. LLMs have demonstrated remarkable performance across various reasoning tasks, even when provided with minimal or no training examples. These models excel in tasks requiring vertical thinking abilities, such as reasoning over physical interactions and social implications (Siino et al., 2022b), showcasing strong common-sense association and inference capabilities. However, prior research has largely overlooked the evaluation of LLMs' lateral thinking abilities, as creative thinking problems are often filtered out during data preprocessing, and only those aligned with common-sense associations are retained.

To address this gap, a novel benchmark called BRAINTEASER (Jiang et al., 2023) to evaluate the lateral thinking abilities of state-of-the-art LLMs is proposed at SemEval-2024 Task 9 (Jiang et al., 2024). The organizers frame lateral thinking puzzles as multiple-choice Question Answering (QA) tasks, a format that is intuitive for humans to engage with and straightforward to assess automatically. The BRAINTEASER benchmark comprises two tasks: Sentence Puzzles and Word Puzzles, designed to assess lateral thinking at different levels of granularity. To develop the dataset, organizers employ a data collection pipeline that retrieves relevant puzzles from publicly available websites, filters out irrelevant question categories, and ensures high data quality. Additionally, to mitigate concerns regarding LLM memorization and consistency, the organizers enhance BRAINTEASER with two reconstruction strategies: semantic reconstruction and context reconstruction. These strate-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 291–297
June 20-21, 2024

gies aim to promote deeper understanding and reasoning rather than mere memorization of patterns.

To meet these objectives, there is a growing demand for automated tools capable of understanding data using recent advancements in NLP models. The emergence of machine and deep learning architectures has sparked increased interest in NLP, prompting substantial efforts to develop techniques for automated identification and understanding of textual content available on the internet. In the literature, various strategies have been proposed so far. Over the past fifteen years, some of the most successful approaches have included Support Vector Machines (SVM) (Colas and Brazdil, 2006; Croce et al., 2022), Convolutional Neural Networks (CNN) (Kim, 2014; Siino et al., 2021), Graph Neural Networks (GNN) (Lomonaco et al., 2022), ensemble models (Miri et al., 2022; Siino et al., 2022), and Transformers (Vaswani et al., 2017).

The sections of this paper are structured as follows: Section 2 offers background information on Task 9, held at SemEval-2024. In Section 3, we outline the approach introduced in this study. Section 4 delves into the specifics of the experimental setup employed to reproduce our findings. The outcomes of the official task and relevant discussions are presented in Section 5. Finally, Section 6 concludes our study and suggests avenues for future research.

We make all the code publicly available and reusable on GitHub[1].

2 Background

The increasing adoption of Transformer-based architectures in academic research has also been bolstered by various methodologies showcased at SemEval 2024. These methodologies tackle diverse tasks and yield noteworthy findings. For instance, at the Task 2 (Jullien et al., 2024), where to address the challenge of identifying the inference relation between a plain language statement and Clinical Trial Reports is used T5 (Siino, 2024c); Task 4 (Dimitrov et al., 2024) where is employed a Mistral 7B model to detect persuasion techniques in memes (Siino, 2024b); and Task 8 (Wang et al., 2024), that utilizes a DistilBERT model to identify machine-generated text (Siino, 2024a).

The Task 9 hosted at SemEval-2024, is based on the human reasoning processes comprising the two already-mentioned types of thinking: vertical and lateral.

Specifically, the BRAINTEASER QA task consists of two subtasks: the Sentence and the Word Puzzle ones, that require awareness of commonsense “defaults” and overwriting them through unconventional thinking that distinguishes these defaults from hard constraints.

In detail, for the Sentence Puzzle one, the puzzle defying common-sense is centred on sentence snippets. On the other hand, for the Word Puzzle subtask, the response diverges from the conventional interpretation of the word and concentrates on the letter arrangement within the target question.

Both subtasks incorporate an adversarial subset, crafted by manually altering the original brainteasers while preserving their underlying reasoning paths.

An example from the official CodaLab page[2] takes as example the following original sentence:

> *"A man shaves everyday, yet keeps his beard long."*

The four possible explanations are:

1. **He is a barber.**
2. He wants to maintain his appearance.
3. He wants his girlfriend to buy him a razor.
4. None of the above.

However, the task organizers also included two other samples based on the previous one. In these two cases, a semantic and a contextual reconstruction have been made to challenge a classification model. The two reconstructions (with the same four possible explanations as in the original) are:

- SEMANTIC RECONSTRUCTION: *"A man preserves a lengthy beard despite shaving every day."*
- CONTEXT RECONSTRUCTION: *"Tom attends class every day but doesn't do any homework."*

3 System Overview

Despite evidence suggesting that Transformers may not always yield optimal results for every text classification task (Siino et al., 2022a), various strategies, such as domain-specific fine-tuning (Sun et al.,

[1] https://github.com/marco-siino/SemEval2024/

[2] https://codalab.lisn.upsaclay.fr/competitions/15566

2019; Van Thin et al., 2023) and data augmentation (Lomonaco et al., 2023; Mangione et al., 2022; Siino et al., 2024a), have proven to be advantageous depending on the task's objectives.

However, to address the task 9 hosted at SemEval-2024 we made use of a zero-shot learning approach (Chen et al., 2023; Wahidur et al., 2024), making use of the DeBERTa Transformer (He et al., 2020).

Our approach is zero-shot (Pourpanah et al., 2022) and make use of the above-mentioned DeBERTa model. Specifically, we employed the multilingual version 3 fine-tuned on the SQuAD2.0 dataset[3]. DeBERTa improves upon the BERT and RoBERTa models by introducing disentangled attention mechanisms and an enhanced mask decoder. Leveraging these enhancements, DeBERTa outperforms RoBERTa across most Natural Language Understanding (NLU) tasks when trained on a dataset of 80GB in size. In DeBERTa V3, efficiency is further enhanced by integrating ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing. Comparative analysis against DeBERTa reveals notable enhancements in model performance across downstream tasks in the V3 version. Further elaboration on the novel techniques employed in this new model can be found in the original paper. The version of DeBERTa utilized in our experiments is mDeBERTa, a multilingual variant of DeBERTa. It maintains an identical architecture while being trained on CC100 multilingual data. The mDeBERTa V3 base model comprises 12 layers with a hidden size of 768. It encompasses 86 million backbone parameters and a vocabulary of 250,000 tokens, resulting in 190 million parameters in the embedding layer. This model underwent training using 2.5 trillion tokens of CC100 data, akin to the XLM-R model.

For the experimental settings, we started evaluating several prompt engineering strategies (White et al., 2023; Liu et al., 2023) to optimize the model replies and to obtain satisfactory results guided by the labelled samples in the training set. For example, we included in the prompt/question to the model, the premise that the given question is a brain-teaser one. Furthermore, we also evaluated the performance of the model on the training set using a few-shot learning setup. In this case, we provided as input (included in the prompt) ten questions indicating the correct answer. Also in this case we did not obtain satisfactory results.

[3]https://rajpurkar.github.io/SQuAD-explorer/

More specifically, given the task hosted at SemEval-2024, we asked the model: *"What is the correct answer to the brain teaser question from the following choices? (Pick only one Option (A)-(D)"*. To this request, the model replied with one or more words that we parsed to extract one of the choices. For example, given the context:

> *"Romeo and Juliet are discovered dead on the bedroom floor. Glass shards and some water were on the floor when they were found. A bookcase and a bed are the sole pieces of furniture in the space. Other than the neighboring railroad track, the house is located in a rural area. How is that even doable? "*

And our question:

> *"What is the correct answer to the brain teaser question from the following choices? (Pick only one Option (A)-(D)"*

And the answers/options:

> (A): They were sleeping and scared by the sound of track.
>
> (B): The rumble of the train moved the shelf which crushed them.
>
> (C): Romeo and Juliet are fish. The rumble of the train knocked the tank off the shelf, it broke and Romeo and Juliet did not survive.
>
> (D): None of above.

The model replied with:

> *"Romeo and Juliet are fish."*

that we mapped into the label *2* corresponding to the third answer. Finally, we collected all the predictions provided on the test set to into a JSON file with required format to submit our predictions.

During our experiments to build our prompt, we also evaluated other LLMs like GPT-Neo and GPT-NeoX (Gao et al., 2020). However, on the labelled training set, we found better performance of DeBERTa in the responses provided. It is also worth notice that we conducted several experiments to find an effective prompt strategy to address the task.

As indicated in a recent investigation by Siino et al. (Siino et al., 2024b), preprocessing does not significantly impact text classification tasks when employing Transformers. Specifically, the optimal combination of preprocessing strategies closely resembles the performance achieved without any preprocessing at all, particularly in the context of Transformer models. Therefore, to maintain a highly efficient and computationally lightweight system, we opted not to apply any preprocessing to the text.

4 Experimental Setup

We implemented our model on Google Colab. The library we used come from Hugging Face and as already mentioned is a multilingual version of DeBERTa[4]. The dataset provided for all the phases are available on the official competition page. We did not perform any additional fine-tuning on the model. To run the experiment, a T4 GPU from Google has been used. After the generation of the predictions, we exported the results on the format required by the organizers. As already mentioned, all of our code is available on GitHub.

5 Results

Participants in Brain Teaser may participate in any or all of the two subtasks. The organizers created two adversarial questions, semantic and context reconstruction, for each brain-teaser (examples can be found on the Task Home Page). The evaluation metrics applied are Instance-based Accuracy and Group-based Accuracy, defined as follows:

- Instance-based Accuracy: Each question, whether original or adversarial, is treated as an individual instance. Accuracy is reported for the original question, its semantic reconstruction, and context reconstruction.
- Group-based Accuracy: Questions and their corresponding adversarial instances are grouped together. A system earns a score of 1 only if it correctly answers all questions within the group. Accuracy is reported for original and semantic reconstruction and original and semantic and context reconstruction.

In Table 1, we present the outcomes derived from our methodology. They are the same results publicly available on the official final ranking shown on the official task page[5]. The results are about the sentence task, given the fact that we did not take part in the word-related task.

DeBERTa	
Original	0.225
Semantic	0.250
Context	0.275
Ori+Sem	0.200
Ori+Sem+Con	0.075
Overall	0.250

Table 1: The method's performance on the test set. In the table are reported the results obtained and shown on the official task page.

Table 2 presents the performance results of the top three teams alongside the results achieved by the final-ranking team, as displayed on the official task page. While our straightforward approach shows potential for enhancement compared to the top-performing models, it is noteworthy that our method required no additional pre-training. Moreover, the computational resources needed to address the task were manageable, utilizing the free online resources provided by Google Colab.

6 Conclusion

This paper introduces the utilization of a DeBERTa model for addressing Task 9 at SemEval-2024. In our submission, we opted for a zero-shot learning approach, leveraging a pre-trained and fine-tuned Transformer model without further adaptation. Through various experiments, we found it advantageous to construct a prompt containing the question for the model. Subsequently, we provided the context, question, and answer candidates as the prompt, prompting the model to discern the correct candidate answer. Despite the task's inherent complexity, as evidenced by the final ranking, there remains ample room for improvement.

Potential alternative methodologies include leveraging the few-shot learning capabilities of the model or exploring alternative models such as GPT and T5. Additionally, integrating additional data or incorporating samples from training and development sets could yield performance enhancements. Further refinements could be achieved through fine-tuning and framing the problem as a text classification task. Moreover, given the promising results

[4]https://huggingface.co/timpal0l/mdeberta-v3-base-squad2

[5]https://codalab.lisn.upsaclay.fr/competitions/15566

TEAM NAME	Original	Semantic	Context	Ori+Sem	Ori+Sem+Con	Overall
abdelhak (1)	1.000	1.000	0.950	1.000	0.950	0.983
lulu13gjdfnglgr (2)	1.000	0.975	0.925	0.975	0.900	0.967
Maxine (3)	0.975	0.975	0.925	0.975	0.900	0.958
wwangbw (31)	0.300	0.175	0.150	0.075	0.025	0.208

Table 2: Comparing performance on the test set. In the table are shown the results obtained by the first three teams and by the last one. In parentheses is reported the position in the official final ranking.

observed across various tasks, the adoption of few-shot learning or data augmentation strategies could also be explored for improved outcomes (Wang et al., 2023; Maia et al., 2024; Siino et al., 2023; Meng et al., 2024; Muftie and Haris, 2023; Tapia-Téllez and Escalante, 2020; Siino and Tinnirello, 2023).

While our straightforward approach demonstrates potential for refinement, it is noteworthy that it required no additional pre-training. Moreover, the computational resources needed to address the task were manageable, utilizing the free online resources provided by Google Colab.

Acknowledgments

We extend our gratitude to the anonymous reviewers for their insightful comments and valuable suggestions, which have significantly enhanced the clarity and presentation of this paper.

References

Shiming Chen, Ziming Hong, Wenjin Hou, Guo-Sen Xie, Yibing Song, Jian Zhao, Xinge You, Shuicheng Yan, and Ling Shao. 2023. Transzero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12844 – 12861.

Fabrice Colas and Pavel Brazdil. 2006. Comparison of svm and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 169–178. Springer.

Daniele Croce, Domenico Garlisi, and Marco Siino. 2022. An SVM ensemble approach to detect irony and stereotype spreaders on twitter. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2426–2432. CEUR-WS.org.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2023. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *arXiv preprint arXiv:2308.10855*.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2024. Semeval-2024 task 9: Brainteaser: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1996–2010, Mexico City, Mexico. Association for Computational Linguistics.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Markus Knauff. 2013. *Space to reason: A spatial theory of human thought*. Mit Press.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Francesco Lomonaco, Gregor Donabauer, and Marco Siino. 2022. COURAGE at checkthat!-2022: Harmful tweet detection using graph neural networks and ELECTRA. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 573–583. CEUR-WS.org.

Francesco Lomonaco, Marco Siino, and Maurizio Tesconi. 2023. Text enrichment with japanese language to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2708–2716. CEUR-WS.org.

Beatriz Matias Santana Maia, Maria Clara Falcão Ribeiro de Assis, Leandro Muniz de Lima, Matheus Becali Rocha, Humberto Giuri Calente, Maria Luiza Armini Correa, Danielle Resende Camisasca, and Renato Antonio Krohling. 2024. Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer. *Expert Systems with Applications*, 241:122418.

Stefano Mangione, Marco Siino, and Giovanni Garbo. 2022. Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2585–2593. CEUR-WS.org.

Zong Meng, Zhaohui Zhang, Yang Guan, Jimeng Li, Lixiao Cao, Meng Zhu, Jingjing Fan, and Fengjie Fan. 2024. A hierarchical transformer-based adaptive metric and joint-learning network for few-shot rolling bearing fault diagnosis. *Measurement Science and Technology*, 35(3).

Mohsen Miri, Mohammad Bagher Dowlatshahi, Amin Hashemi, Marjan Kuchaki Rafsanjani, Brij B Gupta, and W Alhalabi. 2022. Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, 37(12):11319–11341.

Fuad Muftie and Muhammad Haris. 2023. Indobert based data augmentation for indonesian text classification. In *2023 International Conference on Information Technology Research and Innovation, ICITRI 2023*, page 128 – 132.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*.

Marco Siino. 2024a. Badrock at SemEval-2024 Task 8: DistilBERT to Detect Multigenerator, Multidomain and Multilingual Black-Box Machine-Generated Text. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino. 2024b. Mcrock at semeval-2024 task 4: Mistral 7b for multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino. 2024c. T5-medical at semeval-2024 task 2: Using t5 medical embeddings for natural language inference on clinical trial data. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2021. Detection of hate speech spreaders using convolutional neural networks. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2126–2136. CEUR-WS.org.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2022a. Fake news spreaders detection: Sometimes attention is not all you need. *Information*, 13(9):426.

Marco Siino, Marco La Cascia, and Ilenia Tinnirello. 2022b. Mcrock at semeval-2022 task 4: Patronizing and condescending language detection using multichannel cnn, hybrid lstm, distilbert and xlnet. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 409–417. Association for Computational Linguistics.

Marco Siino, Francesco Lomonaco, and Paolo Rosso. 2024a. Backtranslate what you are saying and i will tell who you are. *Expert Systems*, n/a(n/a):e13568.

Marco Siino, Maurizio Tesconi, and Ilenia Tinnirello. 2023. Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2772–2781. CEUR-WS.org.

Marco Siino and Ilenia Tinnirello. 2023. Xlnet with data augmentation to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2763–2771. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2022. T100: A modern classic ensemble to profile irony and stereotype spreaders. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2666–2674. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024b. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

José Medardo Tapia-Téllez and Hugo Jair Escalante. 2020. Data augmentation with transformers for text classification. In *Advances in Computational Intelligence*, pages 247–259, Cham. Springer International Publishing.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rahman S. M. Wahidur, Ishmam Tashdeed, Manjit Kaur, and Heung-No Lee. 2024. Enhancing zero-shot crypto sentiment with fine-tuned language model and prompt engineering. *IEEE Access*, 12:10146 – 10159.

Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. 2023. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. *IEEE Trans. Circuits Syst. Video Technol.*, 33(12):7789–7802.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico, Mexico.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.