BadRock at SemEval-2024 Task 8: DistilBERT to Detect Multigenerator, Multidomain and Multilingual Black-Box Machine-Generated Text

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

Abstract

The rise of Large Language Models (LLMs) has brought about a notable shift, rendering them increasingly ubiquitous and readily accessible. Across diverse platforms such as social media platforms, news outlets, educational platforms, question-answering forums, and even academic domains, there has been a notable surge in machine-generated content. Recent iterations of LLMs, exemplified by models like ChatGPT and GPT-4, exhibit a remarkable ability to produce coherent and contextually relevant responses across a broad spectrum of user inquiries. The fluidity and sophistication of these generated texts position LLMs as compelling candidates for substituting human labour in numerous applications. Nevertheless, this proliferation of machine-generated content has raised apprehensions regarding potential misuse, including the dissemination of misinformation and disruption of educational ecosystems. Given that humans marginally outperform random chance in discerning between machine-generated and human-authored text, there arises a pressing imperative to develop automated systems capable of accurately distinguishing machine-generated text. This pursuit is driven by the overarching objective of curbing the potential misuse of machine-generated content. Our manuscript delineates the approach we adopted for participation in this competition. Specifically, we detail the fine-tuning and the use of a DistilBERT model for classifying each sample in the test set provided. Our submission is able to reach an accuracy equal to 0.754 in place of the worst result obtained at the competition that is equal to 0.231.

1 Introduction

Large language models (LLMs) are increasingly pervasive and readily accessible, leading to a surge in machine-generated content across a multitude of platforms (Fang et al., 2024). LLMs have demonstrated an impressive ability to generate highly fluent responses to diverse user queries. The eloquent nature of these generated texts renders LLMs appealing candidates for replacing human labour across various scenarios. However, this widespread adoption has sparked concerns regarding the potential misuse of such texts, including the dissemination of misinformation in journalistic contexts and disruptions within educational systems (Tang et al., 2023).

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, 2024b); Task 4 (Dimitrov et al., 2024) and Task 10 (Kumar et al., 2024) where is employed a Mistral 7B model to detect persuasion techniques in memes (Siino, 2024a) and to perform Emotion Recognition in Conversation (ERC) within Hindi-English code-mixed conversations respectively (Siino, 2024c).

Despite human evaluators marginally outperforming random chance in distinguishing between machine-generated and human-written text (Mitchell et al., 2023), the need for automatic methods to detect machine-generated content has become increasingly urgent. This necessity prompted the organizers of Task 8 at SemEval-2024 to focus on developing such methods with the aim of mitigating potential misuse.

Previous efforts in detecting machine-generated text have been made. For instance, (Guo et al., 2023) devised methods to discern whether a text was generated by ChatGPT or authored by a human across various domains. However, these endeavours primarily concentrated on the outputs of ChatGPT.

The RuATD Shared Task 2022 tackled artificial

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 239–245

text in Russian, spanning models for paraphrase generation, text simplification, text summarization, and machine translation (Shamardina et al., 2022). However, their emphasis was on models fine-tuned for specific tasks or domains, which differs from the focus of the Task 8. While (Mitchell et al., 2023) detected outputs of various LLMs such as GPT-2, OPT-2.7, Neo-2.7, GPT-J, and NeoX, it's pertinent to note that these models have become obsolete with the advent of GPT-3 and even GPT-4. The Task 8 hosted at SemEval 2024 was built upon the previous work discussed in (Wang et al., 2023b).

To address these objectives, there is an ongoing demand for automated tools capable of extracting and categorizing data, facilitating the classification with recent NLP models. Recent advancements in the machine and deep learning architectures have spurred heightened interest in Natural Language Processing (NLP). Substantial endeavours have been directed towards devising techniques for the automated identification and categorization of textual content accessible on the internet today. In the literature, to perform text classification tasks, several strategies have already been proposed. In the last fifteen years, some of the most successful strategies have been based on SVM (Colas and Brazdil, 2006; Croce et al., 2022), on Convolutional Neural Network (CNN) (Kim, 2014; Siino et al., 2021), on Graph Neural Network (GNN) (Lomonaco et al., 2022), on ensemble models (Miri et al., 2022; Siino et al., 2022) and, recently, on Transformers (Vaswani et al., 2017; Siino et al., 2022b).

Participants in SemEval-2024 Task 8 could compete for three Subtasks better described in the rest of this paper. However, our team participated in the first Subtask only. The first Subtask (i.e., Subtask A) is the Binary Human-Written vs. Machine-Generated Text Classification one: Participants are tasked with determining, based on a given full text, whether it is human-written or machine-generated. There are two tracks for Subtask A: monolingual (only English sources) and multilingual.

The subsequent sections of the paper are structured as follows: Section 2 offers background information on Task 6, 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 task 8 hosted at SemEval-2024 (Wang et al., 2024) is made of three different Subtasks.

- Subtask A. Binary Human-Written vs. Machine-Generated Text Classification: Given a full text, determine whether it is human-written or machine-generated. There are two tracks for Subtask A: monolingual (only English sources) and multilingual.

- Subtask B. Multi-Way Machine-Generated Text Classification: Given a full text, determine who generated it. It can be human-written or generated by a specific language model.

- Subtask C. Human-Machine Mixed Text Detection: Given a mixed text, where the first part is human-written and the second part is machine-generated, determine the boundary, where the change occurs.

An example from the original dataset - better described in (Wang et al., 2023b) - is the one that follows.

Given the prompt:

> *"Write a Wikipedia article with the title Ẅilliam Whitehouse,; the article should at least have 250 words."*

While the original human-written version is available online[2], the output produced by a ChatGPT 3.5 Turbo was:

> *"William Whitehouse was a 19th-century British engineer and inventor who made significant contributions to the field of hydraulics. He was born in 1825 in Birmingham, England, where he grew up and eventually became an apprentice at the Soho Foundry. This apprenticeship instilled in him a passion for steam engines and hydraulics that would shape his future career. After completing*

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

[2] https://en.wikipedia.org/wiki/William_Whitehouse

his apprenticeship, Whitehouse worked for the Birmingham Water Works Company as an engineer. It was during this time that he began experimenting with hydraulic machinery and invented several innovative devices. In 1850, he developed the first hydraulic press, which was able to exert immense amounts of pressure and became widely used in industrial manufacturing processes. He also created hydraulic lifts for raising and lowering heavy loads and hydraulic cranes for lifting and moving heavy objects. Whitehouse's inventions revolutionized manufacturing and construction industries and helped to increase productivity and efficiency. In addition to hydraulic machinery, he also worked on steam engines and locomotives, improving their performance and reliability. He was regarded as an expert in his field and wrote several papers on hydraulic engineering that were widely read and respected. Whitehouse's achievements did not go unnoticed by the scientific community, and he was awarded the Royal Society's prestigious Telford Medal in 1875. He also became a fellow of the Royal Society and was elected to the Institution of Civil Engineers and the Institution of Mechanical Engineers. Despite his success, Whitehouse remained humble and dedicated to creating innovative and practical solutions to engineering problems. He passed away in 1905, leaving behind a legacy of groundbreaking hydraulic inventions that continue to play a vital role in modern manufacturing and construction industries."

3 System Overview

Even if it has already been proved that the Transformers are not necessarily the best option for any text classification task (Siino et al., 2022a), depending on the goal, some strategies like domain-specific fine-tuning (Sun et al., 2019; Van Thin et al., 2023), or data augmentation (Lomonaco et al., 2023; Mangione et al., 2022; Siino et al., 2024a) can be beneficial for the considered task.

However, to address the Task 8 hosted at SemEval-2024 we employed a zero-shot learning strategy (Chen et al., 2023; Wahidur et al., 2024), making use of DistilBERT (Sanh et al., 2020), fine-tuned on the SST-2 dataset (Socher et al., 2013).

DistilBERT, akin to its larger counterparts (i.e., BERT), exhibits commendable performance across a diverse array of tasks when fine-tuned. While prior research predominantly delved into distillation techniques for crafting task-specific models, the distillation approach in this case harnesses knowledge distillation during the pre-training phase. DistilBERT demonstrate the feasibility of reducing the size of a BERT model by 40%, while retaining 97% of its language understanding prowess and achieving 60% increase in speed. To harness the inductive biases inherent in larger models during pre-training, a triple loss mechanism is introduced with this model. This mechanism combines language modelling, distillation, and cosine-distance losses. The compact, expedited, and resource-efficient model not only streamlines the pre-training process but also showcases its potential for on-device computations through a proof-of-concept experiment and comparative on-device analysis.

The Stanford Sentiment Treebank stands as the inaugural corpus equipped with fully labeled parse trees, facilitating comprehensive exploration of the compositional effects of sentiment in language. It comprises 11,855 individual sentences culled from film reviews. Leveraging the Stanford parser, the corpus encompasses a total of 215,154 unique phrases, each annotated by three human evaluators. This novel dataset affords an opportunity to delve into the intricacies of sentiment analysis and capture nuanced linguistic phenomena. Numerous examples within the corpus exhibit distinct compositional structures. The granularity and breadth of this dataset are poised to empower the community in training compositional models grounded in supervised and structured machine learning methodologies. While extant datasets primarily focus on document and chunk labelling, there remains a pressing need to enhance sentiment capture from concise remarks, such as those found in Twitter data.

Utilizing DistilBERT trained on the SST Stanford dataset for detecting human or AI-generated text holds significant promise due to its nuanced understanding of sentiment and context. By leveraging DistilBERT's fine-grained sentiment analysis capabilities, coupled with its proficiency in discerning contextual nuances, the model we used is supposed to effectively distinguish between human-generated and AI-generated text. The SST dataset, annotated for human sentiments classification task, enables DistilBERT to grasp the subtleties of human language, making it adept at identifying deviations indicative of AI-generated content. Moreover, fine-tuning DistilBERT on this dataset enhances its sensitivity to linguistic cues that differentiate human-authored texts from those generated by AI algorithms, thereby offering a robust solution for text authenticity verification in various applications, including misinformation detection, content moderation, and forensic linguistics.

In this study, we employed a fine-tuning approach to enhance the performance of DistilBERT, initially trained on the SST dataset, for the task of distinguishing between human and AI-generated text. The fine-tuning process involved training the model for three epochs on the provided training set, utilizing a portion of the data for validation. Specifically, we partitioned 20% of the training set samples to form a validation set, crucial for assessing the model's performance and preventing overfitting. After completing the fine-tuning process, we systematically evaluated the model's performance across the three epochs on the validation set. Subsequently, we selected the tuned version of the model that exhibited superior performance, as determined by its validation set accuracy. This validation methodology ensures the reliability and generalization capability of the fine-tuned DistilBERT model for the targeted task of differentiating between human and AI-generated text.

In a recent study (Siino et al., 2024b), has been shown that the contribution of preprocessing for text classification tasks is generally not impactful when using Transformers. More specifically, the best combination of preprocessing strategies is not very different from doing no preprocessing at all in the case of Transformers. For these reasons, and to keep our system highly fast and computationally light, we have not performed any preprocessing on the text.

4 Experimental Setup

We implemented our model on Google Colab. The library we used comes from HuggingFace[3] and is the uncased version of DistilBERT specifically trained on the above-mentioned SST2 dataset [4]. We did perform a three-epochs additional fine-tuning, before generating the prediction on the unlabelled test set. This model is versatile and can serve as a foundational tool for topic classification tasks. While it can function as a raw model for masked language modelling or next sentence prediction, its primary utility lies in its adaptability for fine-tuning on downstream tasks. Users can explore the model hub to discover fine-tuned versions tailored for specific tasks beyond its original scope. As already mentioned, all of our code is available on GitHub.

5 Results

Given the binary nature of the classification task, the organizers proposed *Accuracy* as the evaluation metric to be considered for the final ranking. The accuracy is defined in the Equation 1. Where TP stands for the number of correctly predicted right answers, FP stands for the number of wrongly predicted right answers, and FN stands for the right answers wrongly predicted as wrong answers.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In Table 1, we present the outcomes derived from our methodology. They are the same results publicly availble on the official final ranking shown on the official task page[5] and on CodaBench[6].

Compared to the best performing models, our simple approach exhibits some room for improvements. It is worth notice that required no further pre-training and the computational cost to address the fine-tuning stage is manageable with the free online resources offered by Google Colab. However, even with the low effort required, it is possible to achieve interesting results with our proposed approach. Out of the 137 participants, our approach, based on the use of a fine-tuned version of DistilBERT, is able to rank between the position 68 and 69 in the final ranking.

6 Conclusion

This paper presents the application of a DistilBERT-model for addressing the Task 8 at SemEval-2024.

[3] https://huggingface.co/
[4] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english
[5] https://github.com/mbzuai-nlp/SemEval2024-task8
[6] https://www.codabench.org/competitions/1752/

TEAM NAME	Accuracy
safeai (1)	0.969
comp5 (2)	0.961
halwhat (3)	0.961
baseline (19-20)*	0.885
DistilBERT (68-69)*	0.754
saibewaraditya (137)	0.231

Table 1: Comparing performance on the test set. In the table are shown the results obtained by the first three teams, by the last one and by our approach. In parentheses is reported the position in the official final ranking. Our approach is not ranked in the official final ranking, but the score obtained ranks between the positions 68 and 69.

For our submission, we decided to fine-tune a pre-trained Transformer. The model was used to perform a sequence classification task to detect if a piece of text is written by a human or by a generative model. The task is challenging, and there is still opportunity for improvement, as can be noted looking at the final ranking. Possible alternative approaches to our can include utilizing the few-shot capabilities or also the use of other models like Llama and T5, eventually using further data, or directly integrating other samples from the training and from the development sets. Further improvements could be obtained with a fine-tuning and modelling the problem as a text classification task. Furthermore, given the interesting results recently provided on a plethora of tasks, also other few-shot learning (Wang et al., 2023a; Maia et al., 2024; Siino et al., 2023; Meng et al., 2024) or data augmentation strategies (Muftie and Haris, 2023; Tapia-Téllez and Escalante, 2020; Siino and Tinnirello, 2023) could be employed to improve the results. Looking at the final ranking, our simple approach exhibits some room for improvements. However, it is worth notice that it has required no further pre-training and the computational cost to address the task is manageable with the free online resources offered 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.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):1–20.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

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.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

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.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Cherniansk ii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. *arXiv preprint arXiv:2206.01583*.

Marco Siino. 2024a. 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. 2024b. 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. 2024c. Transmistral at semeval-2024 task 10: Using mistral 7b for emotion discovery and reasoning its flip in conversation. 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.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

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.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

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. 2023a. 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, Alham Fikri Aji, and Preslav Nakov. 2023b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

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.