A Prompt in the Right Direction: Prompt Based Classification of Machine-Generated Text Detection

Rinaldo Gagiano and **Lin Tian**
School of Computing Technologies, RMIT University, Melbourne, Australia
{rinaldo.gagiano, lin.tian2}@student.rmit.edu.au

Abstract

The goal of ALTA 2023 Shared Task is to distinguish between human-authored text and synthetic text generated by Large Language Models (LLMs). Given the growing societal concerns surrounding LLMs, this task addresses the urgent need for robust text verification strategies. In this paper, we describe our method, a fine-tuned Falcon-7B model with incorporated label smoothing into the training process. We applied model prompting to samples with lower confidence scores to enhance prediction accuracy. Our model achieved a statistically significant accuracy of 0.991.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has significantly facilitated the generation of complex, human-like text at scale (OpenAI, 2023). These LLMs have found applications in various domains, including AI-assisted writing (Coenen et al., 2021), medical question answering (Yang et al., 2022; Haq et al., 2021, 2022), financial (Lumley, 2023; Haas, 2023; Delocski, 2023), and legal sectors (Trautmann et al., 2022; Blair-Stanek et al., 2023). Leading models like OpenAI's GPT-3 (Brown et al., 2020), Meta's OPT (Zhang et al., 2022), and Big Science's BLOOM (Scao et al., 2022) have the ability to produce content that closely mimics human-created text, making it challenging to distinguish between machine-generated and human-generated content. However, it's important to note that these models lack a genuine understanding of the content they generate.

This limitation can lead to intended negative consequences when this machine-generated content is used in downstream applications. For instance, LLMs have been used to carry out academic fraud (Cotton et al., 2023; Wahle et al., 2022; Elali and Rachid, 2023), disseminate fabricated news stories (Bagdasaryan and Shmatikov, 2022; Groll, 2023; Zellers et al., 2019), and manipulate public opinion (Goldschmidt, 2019; Stella et al., 2018; Bessi and Ferrara, 2016). Given the widespread use of LLMs by the general public (Gault, 2023) and the rapid global dissemination of information, there is a growing risk of disinformation affecting both individuals and organisations.

To address these issues, it is crucial to differentiate between content authored by LLMs and humans. This distinction is essential for ensuring that machine-generated content is used appropriately in various applications while maintaining oversight. Understanding the specific LLM responsible for generating content can help users be aware of potential biases and limitations associated with that model. This interest has led to active research in the area of automatic detection of AI-generated text. Recent work, such as DetectGPT (Mitchell et al., 2023), focuses on techniques for identifying AI-generated content by perturbing text samples and comparing log probabilities. Other approaches involve using LLMs such as DeBERTa (He et al., 2020) or ensemble methods (Przybyła et al., 2023) for multi-class AI detection tasks, illustrating the evolving nature of this research domain.

In this paper, we present our participation in the ALTA 2023 Shared Task (Molla et al., 2023), which centres on the automatic detection of synthetic text produced by LLMs. Participants are challenged with the task of identifying synthetic text across a wide spectrum of sources, spanning different domains and LLMs, including prominent models like T5 (Raffel et al., 2020) and GPT-X (Black et al., 2022). The primary assessment criterion is accuracy, and participants are encouraged to explore diverse methodologies and approaches to construct effective text detection systems.

Our approach involved the fine-tuning of a Falcon-7B (Institute, 2023) model, complemented by the integration of label smoothing during the

training process. Furthermore, we leveraged prompting techniques (Liu et al., 2023) for samples exhibiting lower confidence scores, to guide our model, resulting in improved predictions and an overall enhanced system accuracy.

Our participation in this shared task yielded a successful outcome, as our method attained an overall accuracy of 0.991. This achievement underscores the effectiveness of our approach in discerning between human-authored and LLM-generated text, making a substantial contribution to the ongoing endeavours aimed at addressing the challenges associated with synthetic text.

2 Related Work

Text classification is a field that extensively investigates the extraction of features from unprocessed text data to predict text categories. This topic has witnessed substantial research efforts over recent decades, leading to the development of various models tailored for this purpose.

Traditional models like Naive Bayes, Logistic Regression, Support Vector Machines, Random Forest, and K-Nearest Neighbors have been widely explored (Shah et al., 2020; Pranckevičius and Marcinkevičius, 2017). Machine learning boosting techniques, including Extreme Gradient Boosting and Adaptive Boosting, have demonstrated their prowess in delivering high performance (Stein et al., 2019; Qi, 2020; Tang et al., 2020; Minastireanu and Mesnita, 2019; Bloehdorn and Hotho, 2006). Deep learning models, such as Convolutional Neural Networks and Recurrent Neural Networks, have surpassed traditional methods in text classification tasks (Yogatama et al., 2017; Bharadwaj and Shao, 2019; Zhou et al., 2016).

In recent years, Transformer-based language models have risen to prominence for natural language processing tasks due to their enhanced parallelization capabilities and self-attention mechanisms (Vaswani et al., 2017), compared to prior models like RNNs (Medsker and Jain, 1999). However, it's crucial to acknowledge that while Transformer models excel in the domains for which they were trained, they can be less adaptable when dealing with out-of-domain or unseen samples. Their profound understanding of specific contexts, stemming from vast pre-training data, makes them experts in those domains, yet can hinder their ability to generalise effectively (Gagiano et al., 2021; Sarvazyan et al., 2023; Wang et al., 2023; Li et al., 2023). The focus on the knowledge they acquire during fine-tuning might result in a degree of "domain bias," making them less suitable for broader applications.

To mitigate the limitations of domain-specificity in Transformer models, a hybrid approach in text classification is increasingly gaining recognition (Przybyła et al., 2023; Abburi et al., 2023). The concept of ensembling Transformer models with traditional approaches, such as Naive Bayes, Support Vector Machines, or Ensemble Learning, can harness the benefits of both worlds (Przybyła et al., 2023; Abburi et al., 2023). The specialised domain knowledge acquired by Transformer models can be combined with the interpretability, simplicity, and robustness offered by traditional techniques, ultimately leading to more versatile and adaptive text classification models.

3 Dataset

3.1 Description

The dataset for the ALTA 2023 shared task on binary classification, aimed at distinguishing between human-generated and machine-generated text in English, is sourced from a diverse array of text origins. While not specifically annotated, sources mentioned in the task description encompass various domains, such as law and medicine, and utilise text generated by a range of large language models, including T5 and GPT-X. The dataset has a balanced distribution of human and machine-generated labels, with 9000 samples each, totalling 18,000 samples altogether.

3.2 Pre-processing

In the pre-processing phase, we derive our validation set from the original training data. To achieve this, we initiate the process by tokenising each sample within the training set. Subsequently, we sort these tokenised samples by their respective lengths. When creating subsets from the original training set, we ensure a balanced representation of sample lengths and origin labels. The resulting data splits comprise 15,000 samples for training and 3,000 for validation. This approach facilitates robust model evaluation and ensures that the dataset adequately represents the variations present in the training data.

4 Methodology

4.1 Proposed Approach

In our approach, we used a multi-step strategy to enhance the performance of our text classification task. First, we fine-tuned the Falcon-7B model with label smoothing regularisation on the training data. We then predict on the validation set, obtaining prediction labels and confidence scores. We extract samples below a chosen confidence threshold and use these to prompt our trained model with a predefined prompt. After prompting we predict on the validation set again, using prediction accuracy to determine the optimal confidence threshold.

4.2 Model

Our approach relied on the *Falcon-7B*[1] built by the Technology Innovation Institute[2]. The model is a causal decoder-only model, trained on 1,5000B tokens from the English dataset *RefinedWeb* (Penedo et al., 2023)

4.3 Label Smoothing

Label smoothing is a common regularisation technique in machine learning, especially in neural network training. Large language models often suffer from overconfidence in prediction tasks. To address this issue, label smoothing introduces a small degree of uncertainty, typically controlled by a small value (epsilon, ϵ), into the ground-truth labels during training. Instead of using 1 for the correct class and 0 for all others in classification, label smoothing assigns slightly lower than 1 to the correct class and slightly higher than 0 to the rest. By encouraging the model to acknowledge alternative possibilities and distribute some probability mass to incorrect classes, label smoothing enhances generalisation, making the model more robust and adaptable to unseen data.

4.4 Prompting

Model prompting is a natural language processing technique that transforms the decision-making process of language models. In traditional classification tasks, models analyse entire text inputs and make predictions based on their understanding of the complete content. However, model prompting introduces a novel approach by providing partial inputs or prompts that guide the model's reasoning towards a specific classification. We use the following prompting structure:

"'*{sample_text}*' this is the wrong classified sample, predicted as *{pred_label}* generated with confidence score *{conf_score}* and the gold prediction is *{true_label}*."

This approach significantly influences the model's thinking, rendering it more focused and contextually attuned to the intended classification task.

5 Experiments

5.1 Implementation Details

The parameters we used for model training, label smoothing, and confidence threshold assessment are as follows:

- The hyper-parameters used for model fine-tuning are shown in Table 1.

Parameter	Value
learning_rate	2e-4
fp16	True
max_grad_norm	0.3
max_steps	1000
warmup_ratio	0.03
max_seq_length	512
max_gen_token	1

Table 1: Model fine-tuning hyper-parameters.

- For label smoothing, we set $\epsilon = 0.1$.
- To identify which samples we use for prompting, we search across confidence threshold values of [0.85, 0.92], finding 0.91 optimal.

6 Results

The organisers of the ALTA 2023 shared task provided both a development and a test set for evaluation. While predictions were made on both sets, it's worth noting that the official rankings are determined based on the results from the test set. Accuracy is the metric used to assess the model's performance. For this paper, we exclusively present the results of our test set predictions. The comprehensive leaderboard can be accessed on the ALTA CodaLab Competition website[3].

[1] https://huggingface.co/tiiuae/falcon-7b
[2] https://www.tii.ae/
[3] https://codalab.lisn.upsaclay.fr/competitions/14327

Team Name	Accuracy
OD-21	**0.9910**
DetectorBuilder	0.9845
AAST-NLP	0.9835
SamNLP	0.9820
Organizers	0.9765
VDetect	0.9715
cantnlp	0.9675
ScaLER	0.9665
SynthDetectives	0.9555

Table 2: External evaluation of submissions on the test set. Our approach is highlighted in boldface.

Our approach, under the team name *OD-21*, as showcased in Table 2, achieved the highest accuracy score of 0.9910, as indicated by the boldface. The organisers, using McNemar and Bootstrap tools, determined the result as statistically significant when compared to the closest competing score.

All scores presented in Table 2 are above 0.95. This can be attributed to the favourable circumstances of an in-domain problem. In-domain problems, where the test set originates from the same source as the training data, tend to yield high accuracy, as is evident in our results. This alignment between training and test data contributes to the robust performance of language models in such scenarios.

7 Conclusion

In this paper, we have presented our submission to the ALTA 2023 shared task, a binary classification challenge distinguishing generative AI content from human writing. Our proposed approach, using a Falcon-7B language model combined with label smoothing and model prompting, has demonstrated considerable promise. With a top-ranking accuracy score of 0.991, our system has showcased the effectiveness of these techniques in this specific task. Looking forward, there is an opportunity for further research and refinement. Future work should focus on extending our system's capabilities to tackle more challenging scenarios, including out-of-domain problems and multi-class authorship attribution tasks.

8 Acknowledgments

This research is supported in part by the Defence Science and Technology Group, Australia and the Australian Research Council Discovery Project DP200101441.

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*.

Eugene Bagdasaryan and Vitaly Shmatikov. 2022. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE.

Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7).

Pranav Bharadwaj and Zongru Shao. 2019. Fake news detection with semantic features and text mining. *International Journal on Natural Language Computing (IJNLC) Vol*, 8.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*.

Stephan Bloehdorn and Andreas Hotho. 2006. Boosting for text classification with semantic features. In *Advances in Web Mining and Web Usage Analysis: 6th International Workshop on Knowledge Discovery on the Web, WebKDD 2004, Seattle, WA, USA, August 22-25, 2004, Revised Selected Papers 6*, pages 149–166. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2023. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, pages 1–12.

Boris Delocski. 2023. Natural language processing and its applications in the finance sector.

Faisal R Elali and Leena N Rachid. 2023. Ai-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, 4(3).

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Jenny Zhang, and Jennifer Biggs. 2021. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 119–127.

Matthew Gault. 2023. Ai spam is already flooding the internet and it has an obvious tell.

Cassio Goldschmidt. 2019. Council post: Ai-generated reviews threaten business reputations.

Elias Groll. 2023. Researchers: Large language models will revolutionize digital propaganda campaigns.

Chain Haas. 2023. Introducing bloomberggpt, bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance.

Hasham Ul Haq, Veysel Kocaman, and David Talby. 2021. Deeper clinical document understanding using relation extraction. *arXiv preprint arXiv:2112.13259*.

Hasham Ul Haq, Veysel Kocaman, and David Talby. 2022. Mining adverse drug reactions from unstructured mediums at scale. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 361–375. Springer.

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

Technology Innovation Institute. 2023. Falcon-7b. https://huggingface.co/tiiuae/falcon-7b.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild. *arXiv preprint arXiv:2305.13242*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liz Lumley. 2023. Large language models advance on financial services.

Larry Medsker and Lakhmi C Jain. 1999. *Recurrent neural networks: design and applications*. CRC press.

Elena-Adriana Minastireanu and Gabriela Mesnita. 2019. Light gbm machine learning algorithm to online click fraud detection. *J. Inform. Assur. Cybersecur*, 2019:263928.

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*.

Diego Molla, Haolan Zhan, Xuanli He, and Qiongkai Xu. 2023. Overview of the 2023 alta shared task: Discriminate between human-written and machine-generated text. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association (ALTA 2023)*.

OpenAI. 2023. Gpt-4 technical report.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Tomas Pranckevičius and Virginijus Marcinkevičius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Piotr Przybyła, Nicolau Duran-Silva, and Santiago Egea-Gómez. 2023. I've seen things you machines wouldn't believe: Measuring content predictability to identify automatically-generated text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023). CEUR Workshop Proceedings, CEUR-WS, Jaén, Spain*.

Zhang Qi. 2020. The text classification of theft crime based on tf-idf and xgboost model. In *2020 IEEE International conference on artificial intelligence and computer applications (ICAICA)*, pages 1241–1246. IEEE.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. 2020. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5:1–16.

Roger Alan Stein, Patricia A Jaques, and Joao Francisco Valiati. 2019. An analysis of hierarchical text

classification using word embeddings. *Information Sciences*, 471:216–232.

Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440.

Chaofei Tang, Nurbol Luktarhan, and Yuxin Zhao. 2020. An efficient intrusion detection method based on lightgbm and autoencoder. *Symmetry*, 12(9):1458.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

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.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. How large language models are transforming machine-paraphrased plagiarism. *arXiv preprint arXiv:2210.03568*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. 2017. Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yujun Zhou, Bo Xu, Jiaming Xu, Lei Yang, and Changliang Li. 2016. Compositional recurrent neural networks for chinese short text classification. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 137–144. IEEE.