An Ensemble Based Approach To Detecting LLM-Generated Texts

Ahmed El-Sayed, Omar Nasr
Arab Academy for Science and Technology
{ahmedelsayedhabashy,omarnasr5206}@gmail.com

Abstract

Recent advancements in Large Language models (LLMs) have empowered them to achieve text generation capabilities on par with those of humans. These recent advances paired with the wide availability of those models have made Large Language models adaptable in many domains, from scientific writing to story generation along with many others. This recent rise has made it crucial to develop systems to discriminate between human-authored and synthetic text generated by Large Language models (LLMs). Our proposed system for the ALTA shared task, based on ensembling a number of language models, claimed first place on the development set with an accuracy of 99.35% and third place on the test set with an accuracy of 98.35%.

1 Introduction

In the realm of human-computer interactions, the recent advancements in AI-generated texts are hallmarked by the introduction of Large Language Models (LLMS), such as GPT4 (OpenAI, 2023), GPT3 (Brown et al., 2020), T5 (Raffel et al., 2020), LLAMA (Touvron et al., 2023) and much more. This has resulted in AI's ability to generate text of high quality and fluency comparable to that of humans. These language models have had widespread integration and adaptations across many different fields including but not limited to, law, medicine and education. Nonetheless, similar to any revolutionary technology, LLMs possess both positive and negative aspects for our society. Apart from spreading misleading information, the potential misuse of LLMs could lead to numerous social and ethical challenges, such as academic misconduct (Yun et al., 2023) and spread of misinformation (Else, 2023). The recent growth in adaption of Large Language Models in many domains and their unprecedented ability to generate high quality fluent text similar to that of humans have caught researchers' attention. This lead to the development of systems with the goal of being able to differentiate between human-generated texts and machine-generated ones. Those systems vary according to their scope of operation, ranging from domain specific ones that detect deep fakes based on specific models to more generalized ones, yet there have been efforts to build a unified model able to operate on different domains and generalize to novel LLMs despite not being trained on their respective data. Large Language Models are expected to fundamentally change many aspects of life and with the trend in the number of Large Language Models introduced each year (Naveed et al., 2023), The challenges of detecting text generated by Large Language Models are expected to reach new heights in the upcoming years. The ALTA 2023 shared task (Molla et al., 2023) focuses on this important topic, offering a dataset for evaluation and training. The dataset addresses several issues and supports the creation of a single, readily generalizable model. Our proposed model uses an ensemble-based approach paired with fine-tuning a number of language models. The structure of this research paper will unfold as follows: The related work section will provide an overview of various solutions explored by different researchers in the context of this problem. Subsequently, the data section will detail the properties of the provided dataset and any preprocessing steps undertaken. In the system description section, we will go through the architecture of our proposed model. The results section will then offer a detailed analysis of the outcomes generated by the proposed system, complemented by a comprehensive evaluation of the model's overall performance. Finally, the summary section will synthesize the paper's content, briefly touch on potential future research directions, and consider possible improvements to the model.

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 164–168

2 Related Work

We will touch on the most recent developments in identifying data produced by Large Language models in the section that follows. Because of the widespread use of LLMs and their possible drawbacks, academics have been particularly interested in this area in recent years. Many researchers have proposed systems that use both deep learning techniques and traditional machine learning models. One interesting approach was when (Solaiman et al., 2019) built a logistic regression based detector which made use of TF-IDF unigram and bigram features. The model was trained on GPT-2 outputs and WebText samples and yielded an accuracy up to 97% at 124 million parameters and up to 93% at 1.5 billion parameters. (Fröhling and Zubiaga, 2021) experimented with a number of conventional machine learning approaches, mainly Support Vector Machines, Random Forests and Logistic Regression. In the realm of deep learning, many models were proposed to tackle the problem of AI-generated text, yet many of the proposed systems either focused on specific domains, or they were model specific (Yang et al., 2023; Mitchell et al., 2023). One interesting system was proposed by (Li et al., 2023) which consisted of training 3 detection models; a language model based on Longformer (Beltagy et al., 2020), FastText (Joulin et al., 2016) and GLTR (Gehrmann et al., 2019) and testing the model on multiple settings to ensure its success ranging from domain-specific & model-specific to unseen domains and unseen models. Many studies have also shown that text written by LLM is more objective and less emotional than human-generated text (Webber et al., 2020). Another factor has to do with the fact that LLMs have a condition called hallucinations, which results from the generation of material that is nonsensical or inconsistent (Ji et al., 2023). Something that makes it possible to apply fact-verification procedures. A different strategy is known as "white box detection," where the detector can monitor any unauthorized or suspicious behavior by inserting hidden watermarks into its outputs and having complete access to the target language model (Abdelnabi and Fritz, 2021).

3 Data

The dataset used is the dataset provided in the ALTA 2023 shared task. Below is an illustration of the dataset distribution. The dataset is derived from a number of sources, including several LLMs (e.g., T5, GPT-X) and domain sources (e.g., legal, medical). The labels are AI-generated and Human-generated, represented as 1 and 0 respectively, which formulate a Binary Classification problem. There were 9000 samples in the training set for each of the corresponding labels, spread evenly. Other than the language model-specific preprocessing, no further preprocessing was used.

Dataset	Train	Dev	Test
Texts	18000	2000	2000

Table 1: Data distribution for the task.

4 System Description

In the subsequent section, we will outline our experimentation on the dataset, highlighting the key stages involved in the development of the previously mentioned system.

4.1 Conventional Machine Learning Models

Our approach commenced with word embedding utilizing diverse pretrained word embedding, incorporating padding, and iterative experimentation with various models such as Support Vector Machines and Logistic Regression. While initially productive, these models did not produce satisfactory results. Consequently, we pivoted towards exploring Deep Learning methodologies, focusing primarily on Language Models to enhance the outcomes.

4.2 Language Models

Language models have demonstrated outstanding results on a variety of tasks in recent years. Other researchers have expanded on this accomplishment by creating other models based on BERT (Devlin et al., 2018). Using the dateset we were given, we fine-tuned many BERT-based models. After evaluating the fine-tuning of DistilBERT (Sanh et al., 2019) on our given dataset, achieving an accuracy of 98.5% on the development set, we decided to adopt Roberta (Liu et al., 2019) as our primary model due to its strong performance in similar scenarios (Zhan et al., 2023). Specifically, fine-tuning Roberta resulted in an impressive accuracy of 99.15%. Additionally, XLMRoberta (Conneau et al., 2019) demonstrated a high accuracy of 98.75%, affirming our decision to select Roberta as the foundational model for our development. While experimenting with different hyperparameters for

Roberta, we maintained a consistent accuracy of 99.15%, indicating that higher results were not attainable. Nevertheless, a notable finding was that, despite identical prediction accuracy to our initial model, there were disparities in the predictions. This realization prompted us to implement an ensemble approach.

4.3 Ensembling

An ensemble of machine learning models is a method that combines many different machine learning models, often of different kinds or versions, to enhance robustness, generalization, and predictive performance. By utilizing the combined intelligence of several models, this method outperforms utilizing a single model in terms of prediction accuracy and stability. Our approach involved employing hard voting, a technique where multiple individual models are trained and make predictions on a given dataset. The final prediction is determined through a "voting" mechanism, where each model in the ensemble "votes" for a specific class (in classification tasks). The final output of the ensemble is based on the majority of votes for a particular class or prediction. We experimented with ensembling multiple learners; DistillBERT, XLMRoberta and Roberta Base, then we ensembled multiple Roberta base models. This resulted in the highest performance of the development set. One approach that was only used in an unofficial submission is ensembling Roberta large models, which was found to outperform our previously mentioned models.

4.4 Experiment settings

The training procedure was conducted using the Google Colab platform for training our pipeline, which has 12.68 GB of RAM, a 14.75 GB NVIDIA Tesla T4 GPU, and Python language. We used ktrain's (Maiya, 2020) fit one cycle, which applies a one cycle policy (Smith, 2018). The learning rate was determined via the lr_plot function, which experiments with a range of learning rates and suggests multiple possible learning rates. The parameters set for our experiment are mentioned in Table 2.

Parameter	**Value**
Epochs	10
Learning Rate	Varying
Batch Size	Varying
Max Length	128
Optimizer	AdamW
Loss Function	Binary Cross Entropy

Table 2: Training parameters.

We experimented with 3 different learning rates for Roberta of 1e-5, 2e-5 and 8.675e-6 as well as different batch sizes of 32, 64 and 128.

5 Results

This section examines how well our suggested AAST-NLP system performed in the ALTA-2023 shared task related to the identification of data produced by big language models. Table 3 presents our results, some of which were not evaluated because of submission limit restrictions.

Model Used	**Validation**	**Test**
BASELINE	50.3%	_
DistillBERT	98.5%	_
XLMRoberta	98.75%	_
Roberta BASE	99.15%	98.25%
Ensemble 1	99.3%	98.35%
Ensemble 2	99.35%	98.35%
Ensemble 3	99.3%	98.6%

Table 3: Accuracy of the models on the respective datasets. Ensemble 1 refers to an ensemble of DistilBERT, Roberta and XLMRoberta. Ensemble 2 refers to an ensemble of 3 Roberta-base models. Ensemble 3 refers to an ensemble of 3 Roberta-large models.

Our ensembled models performed the best of the suggested systems, placing first on the test set and third on the development set, suggesting some progress on ensembling multiple learners. Due to computational power constraints, we initially conducted our experiments using Roberta-base. However, after experimenting with Roberta-large, we discovered that when three Roberta-large models were ensembled, they outperformed our top ranking system on the test set.

6 Discussion

The results of these experiments showed that an ensembling-based approach is worth further exploring in the pursuit of a generalized model for classi-

fying synthetic text data generated by LLMS. Some potential further improvements include adding larger models to the ensemble, such as xlm-roberta-XL. Other improvements include supplementing the development data set with more training data such as the one used in (Li et al., 2023). Another approach could be to further tune the hyperparameters of the individual members of the ensemble, which could lead to marginal improvements in the overall performance of the ensemble. Overall, the system has promising implications and, with more research, could prove very fruitful in combating the spread of fake data in the modern world. Addressing this problem is a very pressing matter as this spread of fake synthetic text data could spread far and wide and have catastrophic effects on the journalism industry,the education industry, along with several other industries.

7 Summary

The presented system, utilizing an Ensemble approach through Hard Voting, was thoroughly described in this study. The conducted experiments were comprehensively addressed. Incorporating pretrained language models, along with ensembling, effectively addresses the challenge of identifying text generated by extensive language models, though there remains room for enhancement. Our forthcoming research will concentrate on evaluating our model in analogous settings, utilizing data generated by recently developed Large Language Models across diverse domains to assess its performance. Another compelling avenue for future investigation involves conducting additional experiments with larger language models, particularly emphasizing the adaptation of Roberta-large to our specific problem. This aspect warrants further exploration in subsequent research endeavors.

References

Sahar Abdelnabi and Mario Fritz. 2021. Adversarial watermarking transformer: Towards tracing text provenance with data hiding.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv (Cornell University)*.

T. B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, A. Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv (Cornell University)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv (Cornell University)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*.

Holly Else. 2023. Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944):423.

Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ*, 7:e443.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ROBERTA: A robustly optimized BERT pretraining approach. *arXiv (Cornell University)*.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

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

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)*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models.

OpenAI. 2023. GPT-4 Technical Report. *arXiv (Cornell University)*.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv (Cornell University)*.

Lauren Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. *arXiv (Cornell University)*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv (Cornell University)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Édouard Grave, and Guillaume Lample. 2023. LLAMA: Open and Efficient Foundation Language Models. *arXiv (Cornell University)*.

Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors. 2020. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text.

Hye Sun Yun, Iain J. Marshall, Thomas A Trikalinos, and Byron C. Wallace. 2023. Appraising the potential uses and harms of LLMs for medical systematic reviews. *arXiv (Cornell University)*.

Haolan Zhan, Xiaoqiong He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. 2023. G3Detector: General GPT-Generated Text Detector. *arXiv (Cornell University)*.