Rambunctious Raccoons at SemEval-2023 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection

Advait Deshmukh, Tyler Cranmer, Shrivatsa Mishra
University of Colorado Boulder,
advait.deshmukh@colorado.edu, gecr2427@colorado.edu, shrivatsa.mishra@colorado.edu

Abstract

This paper presents the submission of Rambunctious Raccoons for the SemEval 2023 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection. Our innovative approach employs a transformer-based ensemble method featuring RoBERTa, BERT, XLNet, and ELECTRA. The ensemble outperformed individual models, achieving an accuracy of 87.6% on the dev set, demonstrating an improvement over any single model. Through preprocessing tailored to each transformer, we showcase the collective strength of diverse language models in effectively discerning machine-generated text across varied domains and languages. This research contributes valuable insights into the intricacies of ensemble methods for black-box text detection tasks.

1 Introduction

In recent years, the landscape of natural language processing has been transformed by the widespread emergence of Large Language Models (LLMs) like ChatGPT, Bard, and others. These models have exhibited remarkable advancements in accuracy and credibility, marking a significant evolution in the field. However, accompanying this surge in LLM capabilities is the imperative to address the identification of machine-generated text. In the context of this challenge, we undertook the monolingual Subtask A(Wang et al., 2023). Whether applied to the realms of cheating detection or other critical areas, the need to distinguish between human and machine-generated content presents a novel and pressing challenge for researchers and practitioners alike.

For this task, we opted to employ an ensemble-based method as the main strategy for our system. For this we used 4 models: BERT, RoBERTa, XLNet and ELECTRA. Leveraging the strengths of multiple models, our approach combines the diverse capabilities of individual components to enhance overall performance. This strategic amalgamation not only promotes versatility but also contributes to improved resilience against potential shortcomings of individual models.

Through this task, we have constructed an ensemble model that achieved an impressive accuracy of 87.6% and an F1 score of 0.88 on the development set specifically designated for this task. Notably, our ensemble model's performance surpassed that of a completely distinct model, underscoring the effectiveness of our approach in enhancing classification accuracy and capturing the nuances of the given data. This outcome underscores the potential of our ensemble-based strategy in providing robust solutions for distinguishing between machine-generated and human-generated text.

2 Background

In our task setup, we concentrated on Subtask A, particularly the monolingual subpart. The primary input consisted of textual data, with examples provided in both the test and development sets. For Subtask A, the test set encompassed 119,757 instances, each comprising text and corresponding labels(1: Machine Generated, 0: Human Generated). The development set, serving as a validation subset, comprised 5,000 items. Some examples have been highlighted in Table 1.

Ensemble models have been widely employed across various domains over the past few decades. They facilitate the integration of diverse models, each capable of examining distinct facets that may influence the outcome, thereby capturing multiple characteristics and the inherent structure of the data(Dong et al., 2020). In the realm of Natural Language Processing (NLP), ensemble models have demonstrated considerable efficacy in tasks

Text	Label	Model
Forza Motorsport is a popular racing game that...	1	chatGPT
Buying Virtual Console games for your Nintendo...	1	chatGPT
Windows NT 4.0 was a popular operating ...	1	chatGPT
The authors introduce a semi-supervised method...	0	human
This paper proposes the Neural Graph Machine t...	0	human
The paper proposes a model that aims at learni...	0	human

Table 1: Example Data

such as Spam Detection (Fattahi and Mejri, 2021), Authorship Identification (Abbasi et al., 2022), Sentiment analysis (Chalothom and Ellman, 2015) and Fake News Detection (Elyassami et al., 2022).

3 System Overview

Our system adopts a sophisticated approach to address the challenges of machine-generated text identification, incorporating a diverse ensemble of models and a linear regression component. The ensemble comprises state-of-the-art models, including BERT(Devlin et al., 2018), RoBERTa(Liu et al., 2019), XLNet(Yang et al., 2019), and ELECTRA(Clark et al., 2020), each chosen for its unique strengths in capturing semantic nuances and contextual information.

3.1 Model Ensemble

- BERT (Bidirectional Encoder Representations from Transformers): Known for its bidirectional contextual embeddings, BERT excels in capturing intricate relationships within the input text, enhancing our system's understanding of context and semantics. This model helps serve as our baseline.

- RoBERTa (Robustly optimized BERT approach): Building upon BERT, RoBERTa refines the pre-training process, optimizing performance and robustness. Its inclusion enriches the ensemble's ability to handle various linguistic subtleties.

- XLNet (eXtreme MultiLabelNet): With a focus on context modeling through permutation-based language modeling, XLNet contributes to the ensemble's diversity, addressing challenges posed by intricate sentence structures.

- ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately): ELECTRA is a pre-trained language representation model introduced by Google Research that follows the transformer architecture. It stands out for its unique pre-training approach, which involves training a model to distinguish between "real" and "fake" tokens in a sentence. This method contrasts with traditional masked language modeling used by models like BERT.

3.2 Ensemble Integration

Our system will undergo rigorous testing with three integration methods: Multi-Layer Perceptron (MLP), Logistic Regression, voting, and decision tree. Previous explorations have found that Logistic Regression provides the best results(Nguyen et al., 2023), we are testing on multiple different models, due to difference in the Transformer models present in ensemble. Each method will be evaluated for its ability to harmonize the predictions of individual models, providing insights into their effectiveness and impact on overall system performance.

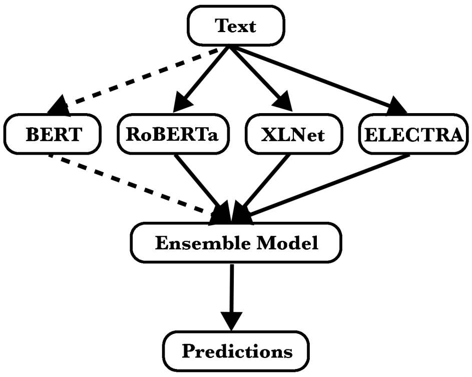

Figure 1: Model Architecture

4 Experimental Setup

4.1 Data Splits

The dataset was partitioned into a test set comprising 119,757 data points and a development (dev) set consisting of 5000 data points. It's noteworthy that only a subset of the test set was utilized for training the models during the fine-tuning of Transformer models. This strategic sampling aimed to mitigate the risk of overfitting and enhance the generalization capability of the models to unseen data. This approach not only contributed to robustness but also facilitated effective hyperparameter tuning and model validation on an independent dataset.

4.2 Preprocessing

In our experimentation, we employed distinct preprocessing pipelines tailored to the requirements of each transformer model. The preprocessing steps encompassed tokenization, padding, and clipping, with careful consideration given to the specific constraints of each model to prevent exceeding maximum input size.

Tokenization: We utilized model-specific tokenizers to convert raw text into tokenized sequences. Each transformer model has its own tokenizer optimized for its architecture and training objectives.

Padding and Clipping: To accommodate variations in sequence lengths and adhere to model constraints, we applied padding to sequences that fell short of the maximum allowable length.
Additionally, we implemented clipping mechanisms to truncate sequences surpassing the model's maximum input size. This ensured compatibility with the diverse requirements of different transformer architectures.

4.3 Hyperparameter Tuning

We have performed Grid Search on both the Logistic Regression and MLP models to optimize their hyperparameters. For Logistic Regression, we explored various combinations of parameters such as regularization strength, solver types, and maximum iterations. Similarly, for the MLP model, hyperparameters including the number of hidden layers, activation functions, and learning rates were systematically tuned.

By tailoring the preprocessing approach for each transformer model, our objective was to optimize the input representations and augment the models' capacity to capture nuanced information from the input text. This fine-tuned preprocessing strategy plays a crucial role in enhancing the overall efficiency and effectiveness of the models, particularly in dealing with diverse linguistic contexts and complexities inherent in the data.

Furthermore, the integration of hyperparameter tuning in the ensemble model further refines the model's configuration. The combined effects of tailored preprocessing and optimized hyperparameters synergistically contribute to the ensemble model's robustness and ability to generalize well across various tasks and linguistic nuances.

5 Results

5.1 Result of individual models

The performance of individual models was noteworthy, with many surpassing the baseline set by RoBERTa for the task. We have also plotted the loss as the model is being trained in Figure 2. Figure 3 presents the performance metrics on the Dev set, showcasing the competence of each model in isolation. As we can see ELECTRA provided the best results on the dev set, with BERT performing the worst. All our models other than BERT have surpassed the baseline provided in the task, having more than 74% accuracy.

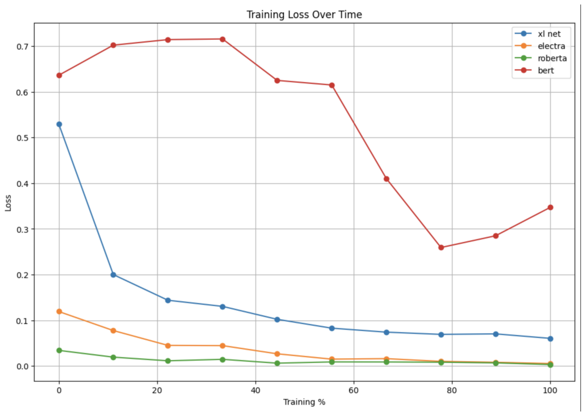

Figure 2: Training Loss Over Time for Different Models

5.2 Ensemble Model

We tried multiple ensemble models, trained on different combinations of each, we found only combining RoBERTA, XLNet and ELECTRA to provide the best results. The inclusion of BERT did not affect the results in any of the models except voting where it increased the F1 score. Upon implementing our ensemble approach, the model exhibited remarkable performance, showcasing an accuracy of 87.6% and an F1 score of 0.88. Notably, these results surpassed the individual model performances,

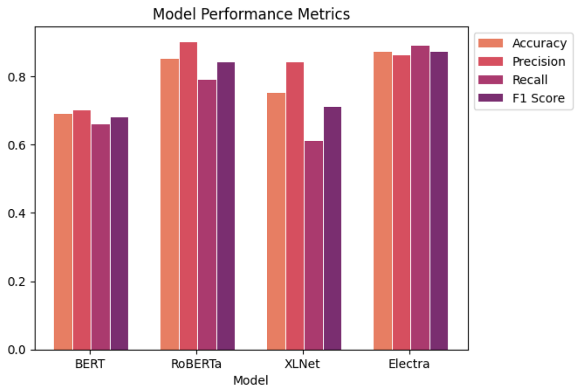

Figure 3: Comparison of Classification Performance Metrics across Models.

soundly outperforming each of them in both accuracy and F1 score metrics.

6 Conclusion

In conclusion, our exploration of ensemble models, while not revealing orders of magnitude improvement, underscores their superiority over individual models across various domains. Notably, the ensemble model consistently outperformed all other models investigated in our study. Surprisingly, our findings highlight that the voting mechanism within the ensemble yielded the most favorable results, surpassing the performance of linear regression and Multi-Layer Perceptron (MLP) models. This unexpected outcome underscores the efficacy of collective decision-making in enhancing predictive accuracy and reinforces the utility of ensemble methods in diverse applications.

Model	Accuracy	F1 Score
BERT	69%	0.68
XLNet	75%	0.71
RoBERTa	85%	0.84
ELECTRA	87.2%	0.87
Decision Tree	85%	0.83
Logistic Regression	86%	0.84
MLP	86%	0.84
Voting Model(Top 4)	86%	0.86
Voting Model(Top 3)	**87.6%**	**0.88**

Table 2: Accuracy and F1 Score of different models on Dev Set

References

Ahmed Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Zunera Jalil, Thippa Reddy Gadekallu, and Natalia Kryvinska. 2022. Authorship identification using ensemble learning. *Scientific reports*, 12(1):9537.

Tawunrat Chalothom and Jeremy Ellman. 2015. Simple approaches of sentiment analysis via ensemble learning. In *information science and applications*, pages 631–639. Springer.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258.

Sanaa Elyassami, Safa Alseiari, Maryam ALZaabi, Anwar Hashem, and Nouf Aljahoori. 2022. Fake news detection using ensemble learning and machine learning algorithms. *Combating Fake News with Computational Intelligence Techniques*, pages 149–162.

Jaouhar Fattahi and Mohamed Mejri. 2021. Spaml: a bimodal ensemble learning spam detector based on nlp techniques. In *2021 IEEE 5th international conference on cryptography, security and privacy (CSP)*, pages 107–112. IEEE.

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 preprint arXiv:1907.11692*.

Duke Nguyen, Khaing Myat Noe Naing, and Aditya Joshi. 2023. Stacking the odds: Transformer-based ensemble for ai-generated text detection. *arXiv preprint arXiv:2310.18906*.

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. 2023. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. *arXiv:2305.14902*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.