Team Innovative at SemEval-2024 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection

Surbhi Sharma
Purdue University
surbhisharma9099@gmail.com

Irfan Mansuri
SWE Qualcomm
iffyaiyan@gmail.com

Abstract

With the widespread adoption of large language models (LLMs), such as ChatGPT and GPT-4, in various domains, concerns regarding their potential misuse, including spreading misinformation and disrupting education, have escalated. The need to discern between human-generated and machine-generated text has become increasingly crucial. This paper addresses the challenge of automatic text classification with a focus on distinguishing between human-written and machine-generated text. Leveraging the robust capabilities of the RoBERTa model, we propose an approach for text classification, termed as RoBERTa hybrid, which involves fine-tuning the pre-trained Roberta model coupled with additional dense layers and softmax activation for authorship attribution. In this paper, we present an approach that leverages Fabien et al. (2020) Stylometric features, hybrid features, and the output probabilities of a fine-tuned RoBERTa model. Our method achieves a test accuracy of 73% and a validation accuracy of 89%, demonstrating promising advancements in the field of machine-generated text detection. These results mark significant progress in the domain of machine-generated text detection, as evidenced by our 74th position on the leaderboard for Subtask-A of SemEval-2024 Task 8.

1 Introduction

SemEval-2024 Task 8 Wang et al. (2024) is centered on the detection of machine-generated text across multiple generators, domains, and languages. This detection is crucial for mitigating the risks associated with the potential misuse of large language models (LLMs), which have advanced capabilities in generating multilingual human-like texts. In this task, the goal is to differentiate between machine-generated and human-authored texts, addressing concerns regarding the authenticity and trustworthiness of textual content in various contexts and languages.

The rapid advancement of deep learning technologies has ushered in a new era where the boundaries between human-generated and machine-generated artifacts are increasingly blurred. This evolution is epitomized by the emergence of Deepfakes, which convincingly mimic genuine human actions, and the widespread adoption of Natural Language Generation (NLG) systems, particularly those leveraging neural language models. These developments have led to the creation of neural texts that bear striking resemblances to human-authored content, posing significant challenges in distinguishing between the two.

Traditionally, Authorship Attribution Uchendu et al. (2020) within the realm of Natural Language Processing (NLP) focused on accurately attributing text to its true human author. However, with the advent of Neural Language Generation (NLG) techniques Uchendu et al. (2023) capable of producing human-quality open-ended texts, the attribution landscape has expanded to encompass authorship by humans, machines, or a combination thereof. As the quality of machine-generated texts continues to improve, the lines between human and machine-generated text become increasingly indistinct, exacerbating the challenge of differentiation.

Moreover, the potential for misuse of these technologies, including the generation of misinformation, fake reviews, and political propaganda at scale Uchendu et al. (2023), underscores the critical need for effective methods to discern neural texts from human-authored content—a problem known as Neural Text Detection (NTD) Uchendu et al. (2023), which is a sub-problem of the broader authorship attribution domain.

In this paper, we present an approach named RoBERTa hybrid, tailored to address the challenge of distinguishing between human and machine-generated texts. This method utilizes the fine-tuning of a pre-trained RoBERTa language model, enhanced with additional layers for text classifica-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1172–1176

tion, in evaluating the performance of pre-trained language models for text differentiation tasks. We specifically target the task of automated detection of machine-generated text, recognizing the growing importance of discerning between human-authored and artificially generated content in today's digital landscape.

In this paper, we present an approach named RoBERTa hybrid, tailored to address the challenge of distinguishing between human and machine-generated texts. This method utilizes the fine-tuning of a pre-trained RoBERTa language model, enhanced with additional layers for text classification, in evaluating the performance of pre-trained language models for text differentiation tasks. We specifically target the task of automated detection of machine-generated text, recognizing the growing importance of discerning between human-authored and artificially generated content in today's digital landscape. We secured the rank 74 on the leaderboard for SubTasK-A

2 Related Work And Background

Subtask A of Task-8 in the SemEval challenge Wang et al. (2024) utilized a monolingual dataset, focusing on distinguishing between human-written and machine-generated text. The dataset primarily employed English as its medium. Each instance in the dataset is labeled as either 1 (indicating machine-generated text, specifically generated by Chat-GPT) or 0 (indicating human-written text).

The Subtask A dataset consists of three subsets: training, development, and testing datasets. The training dataset contains 119,757 samples, while the development dataset comprises 5,000 samples. The testing dataset includes 34,272 samples. In the training dataset, there are 63,351 samples labeled as class 0 and 56,406 samples labeled as class 1. Conversely, the development dataset consists of 2,500 samples, all of which are labeled.

Within the dataset, there are columns denoted as "model" and "source." The "model" column specifies which model generated a particular text, while the "source" column indicates the origin of the text.

Previous research has explored various approaches to authorship attribution (AA), aiming to accurately identify the authors of texts, which is crucial in fields such as forensic linguistics, plagiarism detection, and content analysis. Traditional methods have relied on Stylometric features, which capture the distinctive writing style of individual authors based on linguistic patterns and characteristics. However, recent advancements in natural language processing (NLP) have introduced transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) that have demonstrated state-of-the-art performance across a range of NLP tasks. Fabien et al. (2020) showcased the effectiveness of BERT for text classification tasks, highlighting its ability to extract semantic and syntactic information from text. However, there has been a lack of systematic exploration into the performance of fine-tuned pre-trained language models, specifically for authorship attribution. The introduction of BertAA addresses this gap by fine-tuning BERT with a dense layer and softmax activation specifically for authorship attribution. The incorporation of BERT allows for the utilization of semantic and syntactic information encoded in text representations, potentially improving the accuracy of authorship attribution systems. Furthermore, BertAA integrates Stylometric features, which capture lexical and structural characteristics of text, and hybrid features, which combine character-level n-grams, enhancing the model's ability to capture both content-related and stylistic aspects of authorship.

However, Stylometric classifiers encounter challenges when tasked with accurately determining the authorship of human versus neural texts. Uchendu et al. (2023) noted that certain Stylometric classifiers were surpassed in performance by deep learning-based models. Furthermore, research findings, as cited in Schuster et al. (2020), revealing Stylometry's inability to identify neural misinformation underscore the necessity for alternative methodologies to address the Authorship Attribution (AA) task within the context of Neural Text Generation (NTD). Consequently, researchers have increasingly embraced and refined deep learning-based approaches for distinguishing between neural and human-generated text. These approaches can be further classified into three main categories: Glove-based, Energy-based, and Transformer-based Attribution models.

Language models often exhibit a lack of syntactic and lexical diversity, characterized by the repetition of the same expressions and a limited use of synonyms and references. This behavior Fröhling and Zubiaga (2021) can be approximated using named entities (NE) and properties of coreference chains, along with shifts in part-

of-speech (POS) distributions between human and machine-generated text. Features based on NE-tags, coreference chains, and POS distributions Fröhling and Zubiaga (2021) can effectively capture the differences in syntactic and lexical diversity Gehrmann et al. (2019) between human and machine-generated text.

Repetitiveness: Machine-generated text is prone to repetitiveness Holtzman et al. (2019), often overusing frequent words and exhibiting highly parallel sentence structures. Features such as the share of stop-words, unique words, and words from "top-lists" can highlight the lack of diversity in machine-generated text. Additionally, measures of n-gram overlap in consecutive sentences can reveal patterns of lexical and syntactic repetition, further distinguishing between human and machine-generated text.

Lack of Coherence: A significant challenge in machine-generated text is the lack of coherence Holtzman et al. (2019), particularly over longer sentences and paragraphs. Coherence can be assessed through the development of entities and the tracking of their appearance and grammatical roles across the text. Features based on entity grids and transition frequencies Badaskar et al. (2008) between consecutive sentences can capture the coherence or lack thereof in machine-generated text.

By incorporating these features into automated detection methods, researchers aim to develop robust and accessible tools for distinguishing between human and machine-generated text, thereby mitigating the risks associated with language model abuse.

3 System Overview

The experiments conducted encompassed Subtask A within the monolingual track. Subtask A posed a binary classification challenge, aimed at discriminating between human-generated text and text produced by the Machine (ChatGPT).

In addressing Subtask A, a Stylometric classifier was developed to exploit diverse stylistic attributes, encompassing text length, word count, average word length, count of short words, proportion of digits and capital letters, frequencies of individual characters and digits, hapax-legomena (a measure of text richness), and the frequency of 12 punctuation marks. These Stylometric features were employed in training a Logistic Regression model.

Furthermore, hybrid features, incorporating the 100 most frequent character-level bi-grams and tri-grams, were integrated. Logistic Regression was applied for classification using these hybrid features as well.

The ultimate model adopted a hybrid strategy, whereby output probabilities from the RoBERTa classifier, the Stylometric classifier, and the hybrid features classifier were concatenated. This concatenated output underwent classification using an additional Logistic Regression model. We chose RoBERTa due to its robust performance in natural language understanding tasks, its ability to handle a wide range of text data, and its pre-training on large-scale corpora, which helps capture nuanced linguistic patterns.

To refine the RoBERTa hybrid model, class probabilities derived from the Stylometric features and those obtained from fine-tuning the RoBERTa model were concatenated separately for both the training and test datasets. Additionally, the probabilities derived from training a Logistic Regression model on hybrid features were integrated into the hybrid model.

3.1 Stylometric Features Extraction

- Length of text: Count the number of characters or tokens.
- Number of words: Total word count.
- Average word length: Average length of words.
- Number of short words: Count of words below a certain threshold.
- Proportion of digits and capital letters: Ratio of digits and capitals to total characters.
- Individual letter and digit frequencies: Count of each letter and digit.
- Hapax-legomena: Words occurring only once.
- Frequency of punctuation marks: Count of specific punctuation marks.

3.2 Hybrid Features Extraction

- Character-level n-grams: Extract the 100 top frequent character-level bi-grams and tri-grams in the text.

3.3 Logistic Regression Model

- In logistic regression, the input features are linearly combined, and the result is passed through the logistic function (also known as

the sigmoid function) to obtain the probability of the positive class.

- Mathematically, the logistic regression model can be represented as:

$$P(y = 1|x) = sigmoid(w^T \cdot x + b)$$

where $P(y = 1|x)$ is the probability of the positive class given the input features x, w represents the weight vector, b is the bias term, and $sigmoid$ is the logistic function.

Hyperparameters:

- **Penalty**: This hyperparameter controls the regularization strength, with options typically including L1 (Lasso) or L2 (Ridge) regularization.
- **Tolerance**: It determines the stopping criteria for the optimization algorithm, specifying the tolerance for the change in the loss function between iterations.
- **Maximum Iterations**: This sets the maximum number of iterations allowed for the optimization algorithm to converge.
- **Intercept**: A boolean parameter indicating whether to include an intercept term in the model.

These hyperparameters are crucial for controlling the model's complexity, preventing overfitting, and optimizing performance. They are typically tuned using techniques like grid search or cross-validation to find the best combination for the given dataset and task.

Model	Parameter	Value
Hybrid feat.	Char. N-grams	(2,3)
LR	Penalty	l2
	Tolerance	0.0001
	Cost	1.0
	Max Iterations	100
	Intercept	True
RoBERTa	Max Iterations	100
	Intercept	True
	Config Epochs	1 to 5
	Input token length	512

Table 1: Parameters of the experiments.

4 Experimental Setup and Evaluation Results

Experimental Design: In our experimental setup, we fine-tuned the RoBERTa model over 5 epochs using the training dataset. To ensure model robustness and prevent overfitting, we utilized a validation dataset consisting of 80% of the training data and 20% of the testing data.

Our approach involved creating a hybrid model, which integrated class probabilities from three classifiers: Stylistic classifier, hybrid classifier, and RoBERTa classifier. These probabilities were concatenated and passed through a Logistic Regression layer for training.

To assess the efficacy of our model, we evaluated its performance using the accuracy metric.

Classifier	Accuracy (%)
Stylometric classifier	49
Hybrid Features Classifier	58
RoBERTa + Style Classifier	73
Hybrid Classifier (RoBERTa + Style + Hybrid)	73

Table 2: Accuracy Results on the Test Dataset

Evaluation of Results: The accuracy results on the test dataset are summarized in Table 2. We observe varying performance among different classifiers. The Hybrid Classifier (RoBERTa + Style + Hybrid) achieved the highest accuracy of 73%, outperforming both the Stylometriv Classifier (49%) and the Hybrid Features Classifier 58%. However, (RoBERTa + Style) based Classifier too resulted in 73% This indicates that incorporating RoBERTa-based representations along with Stylometric and hybrid features significantly improved the model's ability to classify text accurately.

The superior performance of the Hybrid Classifier can be attributed to its utilization of RoBERTa, a transformer-based model known for its ability to capture rich contextual information from text. By leveraging RoBERTa's representations along with Stylometric and hybrid features, the Hybrid Classifier achieved a more comprehensive understanding of the input text, leading to better classification accuracy.

On the other hand, the Stylometric Classifier and the Hybrid Features Classifier exhibited lower

accuracies compared to the Hybrid Classifier. This could be due to their reliance on a narrower set of features for classification, which may not capture the full complexity of the input text.

5 Conclusion

In this paper, we have presented an innovative approach, termed the RoBERTa hybrid model, for the task of detecting machine-generated text. Leveraging the robust capabilities of the Roberta model, we fine-tuned it coupled with additional dense layers and softmax activation for authorship attribution. Our method incorporates a hybrid of Stylometric features, character-level n-grams, and the output probabilities of a fine-tuned Roberta model, achieving significant advancements in machine-generated text detection.

Experimental results demonstrate the effectiveness of our approach, with a validation accuracy of 89% and a test accuracy of 73%. Although these results do not surpass the baseline methods, they highlight the potential of our approach in addressing the challenges posed by machine-generated text across diverse domains.

Moving forward, our work opens avenues for further research in enhancing the accuracy and robustness of machine-generated text detection systems. Future efforts may focus on exploring additional feature representations, optimizing model architectures, and addressing the challenges posed by monolingual machine-generated text. Our efforts in enhancing machine-generated text detection we have tried to contribute to the broader objective of safeguarding the integrity and credibility of online content.

Acknowledgements

We extend our heartfelt appreciation to all contributors to this research. Our Supervisor has provided invaluable insights and feedback, enriching the quality of our work. We extend special gratitude to the reviewers for their constructive comments and suggestions, which greatly enhanced the paper. The first author led the experimental design and paper writing process, while the second author with my guidance has contributed exclusively to proofreading and paper writing, not involved in the experimental design.

References

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

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

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 25(1):1–18.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, 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.