UMUTeam at SemEval-2024 Task 8: Combining Transformers and Syntax Features for Machine-Generated Text Detection

Ronghao Pan[1], **José Antonio García-Díaz**[1],
Pedro José Vivancos-Vicente[2], **Rafael Valencia-García**[1]
[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
[2]VÓCALI Sistemas Inteligentes S.L., Parque Científico de Murcia,
Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, España
{ronghao.pan, joseantonio.garcia8, valencia}@um.es
pedro.vivancos@vocali.net

Abstract

These working notes describe the UMUTeam's participation in Task 8 of SemEval-2024 entitled "Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection". This shared task aims at identifying machine-generated text in order to mitigate its potential misuse. This shared task is divided into three subtasks: Subtask A, a binary classification task to determine whether a given full-text was written by a human or generated by a machine; Subtask B, a multi-class classification problem to determine, given a full-text, who generated it. It can be written by a human or generated by a specific language model; and Subtask C, mixed human-machine text recognition. We participated in Subtask B, using an approach based on fine-tuning a pre-trained model, such as RoBERTa, combined with syntactic features of the texts. Our system placed 23rd out of a total of 77 participants, with a score of 75.350%, outperforming the baseline.

1 Introduction

In the area of Natural Language Generation (NLG), advances such as GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022) have provided support for various writing tasks. The widespread adoption of Large Language Models (LLMs) such as ChatGPT and GPT-4 (Achiam et al., 2023) has led to an increase in machine-generated content across various platforms, including news, social media, education and science. While these models produce remarkably fluid responses, concerns have arisen about their potential to spread misinformation and disrupt established systems. Concerns remain about their misuse, particularly in scenarios such as academic dishonesty and scientific research, where AI-generated content may be presented as original work. The emergence of AI-generated scientific texts raises ethical and integrity concerns in academic publishing, requiring tools or models to distinguish between human-generated and AI-generated content (Ma et al., 2023).

Efforts to detect AI-generated text have primarily involved fine-tuning pre-trained models and developing detection systems. Recent studies have presented datasets and methods specifically designed for the detection of AI-generated scientific documents. However, challenges remain in achieving high performance and interpretability across different domains and models (Ma et al., 2023).

For this reason, Task 8 of SemEval (Wang et al., 2024), entitled "Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection", aims at identifying automatic systems for the detection of machine-generated text in order to mitigate its potential misuse. To this end, the task is divided into three subtasks that address two text generation paradigms: (1) full text, where a text is considered to be entirely written by a human or generated by a machine; and (2) mixed text, where a machine-generated text is refined by a human, or a text written by a human is paraphrased by a machine.

This shared task is divided into three subtasks:

- **Subtask A: Binary Human-Written vs. Machine-Generated Text Classification**. Determine whether a given full-text was authored by a human or generated by a machine. It offers two tracks: monolingual (English source only) 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 containing both human-generated and machine-generated

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 697–702

segments, identify the boundary where the transition from human-generated to machine-generated content occurs.

In this competition, the UMUTeam participated only in the **Subtask B** with an approach based on fine-tuning a pre-trained model such as RoBERTa combined with syntactic features of the text. Syntax features of the text refer to the writing style, such as token-level features (e.g. word length, part of speech, function word frequency and stop word ratio) and sentence-level features (e.g. sentence length).

During our experiments, we found that the syntactic features of texts can complement and improve the performance of pre-trained Transformer-based models and that RoBERTa is more suitable for this type of task.

The rest of this paper is organized as follows. First, Section 2 provides a summary of important details about the shared task setup. Second, Section 3 gives an overview of our system. Section 4 presents the specific details of our systems. Section 5 discusses the results of the experiments, and finally, the conclusions are presented in section 6.

2 Background

Recent advances in AI technology, particularly in the field of Natural Language Processing (NLP), have led to the emergence of many models capable of generating natural language using LLMs. These can produce remarkably fluent responses, and this has led to an increase in machine-generated content across multiple domains and platforms, including news, social media, education, and science.

LLMs face several technical and social challenges as they advance in NLP tasks. Recent research has shown that pre-trained LLMs can not only learn linguistic knowledge, but also reason about large amounts of acquired knowledge (Lewis et al., 2020). However, LLMs have other problems, such as hallucination, producing texts that contain information or details that are not based on reality or are completely invented; and asserting falsehoods as facts, which means that they can involuntarily produce texts that present false information as true.

The latest generative LLMs, such as GPT-3, are capable of producing highly fluent text, but they can produce inaccurate, toxic or unhelpful content. Some researchers have explored the use of reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) to adjust language models to better match user intent. ChatGPT, one of OpenAI's models based on GPT-3 and trained with RLHF, performs well in conversations with humans, demonstrating an understanding of user instructions and generating useful, reliable, honest and harmless text content.

Therefore, a growing number of studies have been conducted to analyze, recognize and identify text generated by AI, especially text generated by GPT. Current research focuses on two main areas: human behavior for recognizing text generated by AI and recognition models for identifying text generated by LLMs. For example, in (Guo et al., 2023), an approach was proposed to determine whether a text (in English and Chinese) was generated by ChatGPT or written by a human across different domains, while in (Shijaku and Canhasi, 2023), a model was developed to identify whether TOEFL essays were written by humans or generated by ChatGPT on a small dataset (126 essays for each).

There are other studies that focus on detecting fake information or fake news generated by LLMs. For example, in (Zellers et al., 2019), the Grover model was proposed to generate and detect examples of fake news. After the release of GPT-2, OpenAI proposed the GPT-2 generated text detector, which achieved a high F1 score. This detector was fine-tuned based on RoBERTa in a binary text classification format. In addition, many studies also use various data augmentation techniques to improve model performance in the classification task through external data that complements the model or simply increases the training set (Bayer et al., 2022). In paper (Ma et al., 2023), an approach was proposed to detect text generated by language models using different text features such as writing style, coherence, consistency, and argument logistics. The model with only syntax features (writing style) achieved the best result.

For this shared task, we used a fine-tuning approach of transformer-based models such as RoBERTa to create a detector for text generated by different LLMs. Unlike other existing studies on LLM-generated text detection, we have concatenated syntactic features during the fine-tuning process to improve its performance. The model evaluated for Subtask B is **RoBERTa** (Liu et al., 2019), a model based on Transformers, which was pre-trained on a large corpus of English data with Masked Language Model (MLM) goal. For this task, we evaluated the *base* version.

3 System overview

Figure 1 shows the architecture of our system. First, we extracted the syntactic features of the texts using the syntactic feature extractor and encoded the texts into a vector containing the dense representation of all the information contained in the text by the pre-trained models, i.e., the last hidden state of the model with text as input. Second, once the vector and syntactic features were obtained, we normalized the syntactic feature values and concatenated them with the text vector. Third, the fine-tuning process is performed, and a sequence classification layer is added on top of the pre-trained model. This layer takes the sequence representation generated by the pre-trained model and performs a classification based on the labels of the specific classification task. Finally, a performance evaluation is performed using the validation set.

3.1 Syntactic feature extractor

Syntactic linguistic features are those aspects related to the grammatical structure and organization of words in a sentence or paragraph (García-Díaz et al., 2022b). This can include elements such as sentence length, the frequency of certain parts of speech, the presence of function words, the number of stop words, etc. All of these features reflect the writing style that distinguishes different texts. In general, syntactic linguistic features have proven effective in NLP tasks such as author analysis (García-Díaz et al., 2022a) or hate speech identification (García-Díaz et al., 2023b).

The features used in this task are:

- **Average word length**. This is the average number of characters the words in the text have. It is calculated by adding the length of all words in the text and dividing that sum by the total number of words in the text.
- **POS tag frequency**. This is the frequency of Part of Speech (POS) grammatical tags in the text. Grammatical tags represent the grammatical categories of words in a text, such as nouns, verbs, adjectives, and so on. The frequency of POS tags indicates how often different grammatical categories occur in the text and can provide information about the structure and style of the text.
- **Average sentence length chars**. This is the average length of the sentences in the text, measured in characters. It is a measure of the complexity and readability of the text.
- **Average sentence length words**. This is the average length of the sentences in a text, measured in words. This metric shows the average number of words per sentence in the text.
- **Percentage of stopwords**. This is the percentage of stopwords in the text, relative to the total number of words in the text. Stopwords are common words that are often filtered out or eliminated during natural language processing because they occur so frequently and have little contextual meaning.
- **Punctuation Frequency**. Refers to the number of times that different punctuation marks, such as commas, periods, semicolons, etc., occur in a given text.
- **Special character Frequency**. Refers to the number of times characters other than letters and numbers occur in a given text. These special characters can include punctuation marks, mathematical symbols, control characters, emoticons, and other non-alphabetic symbols.

For the syntactic features, we used an open source tool called *authorstyle*[1], a package that allows to handle digital text forensics and stylometric corpora to extract stylometric features.

The embeddings of the texts refer to the numerical representation of the words or tokens in a high-dimensional vector space obtained by the tokenizers of the models. For this task, we normalized the syntactic feature values and concatenated them with the embeddings obtained by the tokenizers to perform RoBERTa fine-tuning to identify the author. It can be written by humans or generated by a specific language model.

4 Experimental setup

For Subtask B, we used the dataset provided by the organizers, which consists of two subsets: training and validation. Figure 2 shows the distribution of the training and validation sets. We can see that both the training and validation sets are balanced and that there are a total of 5 types of texts generated by different LLMs or by humans. The types

[1] https://github.com/mullerpeter/authorstyle

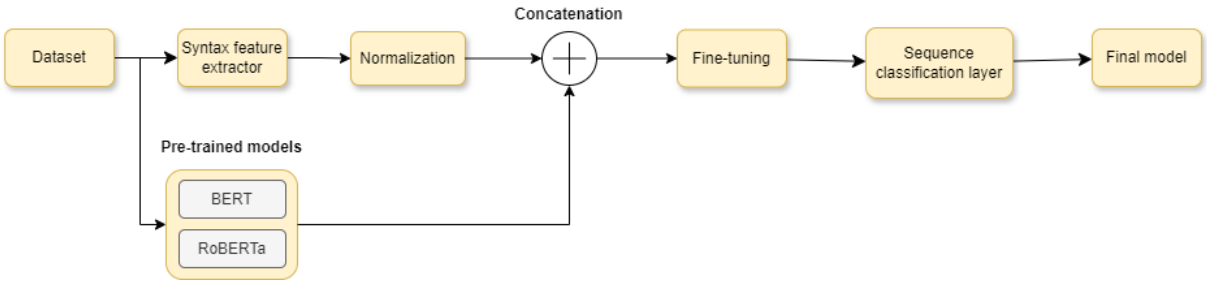

Figure 1: System architecture

are: davinci, bloomz, human, chatGPT, dolly and cohere.

We used the following fine-tuning hyperparameters: a batch size of 16 for both training and validation, 10 epochs, a learning rate of 2e-5, and a weight decay of 0.01.

During training, we used Macro-F1 as a reference. Macro-F1 is a measure used to evaluate the performance of a model in a multi-class classification problem. It calculates the average F1 score for each class individually, and then averages these scores to obtain an overall score. The macro F1 Score assigns equal weight to each class, regardless of its size or distribution in the data set. This means that all classes are equally important in the final scoring metric.

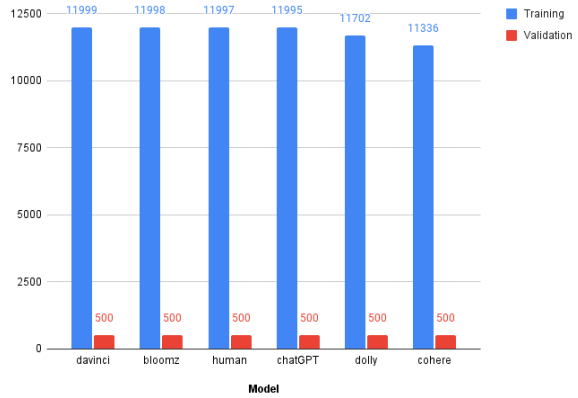

Figure 2: Training and validation set distribution of Subtask B.

5 Results

In the Table 1 we can see the official ranking of Subtask B. With a total of 77 contestants, we have achieved the twelfth-third best result, with an accuracy of 75.350, which is 0.744% higher than the baseline and 15.5% lower than the first.

In order to perform an error analysis and to observe the behavior of our model in predicting different classes of texts, we have generated the confusion matrix for our model based on the test set, as shown in Figure 3. Our analysis shows that our model has a strong predictive performance for texts generated by Bloomz, Dolly, ChatGPT and Davinci, reaching accuracies above 90%. However, when it comes to human-generated texts, our model shows a 27.73% tendency to misclassify them as generated by the Dolly model. In particular, when predicting texts generated by Cohere, our model tends to misclassify them as generated by Davinci at a rate of 70%, leading to a decrease in overall accuracy.

Table 1: Official results for the Subtask B.

Team	Rank	Accuracy
joeblack	1	90.850
tmarchitan	2	86.955
farawayxxc	3	84.328
halwhat	4	83.955
dianchi	5	83.478
	...	
UMUTeam	23	**75.350**
	...	
Baseline	-	74.606

6 Conclusion

This paper describes the participation of the UMUTeam in the 8th shared task of SemEval 2024, focused on the identification of automatic systems for the recognition of machine-generated text in order to mitigate its potential misuse. The task consisted of three subtasks: Subtask A, a binary classification task to determine whether a given full-text was written by a human or generated by a machine; Subtask B, a multi-class classification problem to determine, given a full-text, who gen-

Figure 3: The confusion matrix of our RoBERTa-based system in the test set.

erated it. It can be written by a human or generated by a specific language model; and Subtask C, mixed human-machine text recognition. In this shared task, we participated in Subtask B, using a approach based on fine-tuning a RoBERTa pretrained model with syntactic features of texts. In terms of results, our system achieved the 23rd position with a score of 75.350%, outperforming the baseline.

Due to our line of research, we will evaluate our system on texts containing figurative language (García-Díaz and Valencia-García, 2022) and financial language (García-Díaz et al., 2023a). On the one hand, the ambiguity and creativity of figurative language poses a challenge to the recognition of automatically generated text, as LLMs may have difficulty replicating the creative nuances of human-generated content. On the other hand, the recognition of automatically generated financial and business text is challenging due to specialized vocabulary and complex technical concepts. Ideally, LLMs must have deep domain-specific understanding to produce accurate content that requires regulatory compliance and accuracy, which requires careful review and validation against authoritative sources.

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way to make Europe and LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, this work was funded by the Spanish Government, the Spanish Ministry of Economy and Digital Transformation through the Digital Transformation through the "Recovery, Transformation and Resilience Plan" and also funded by the European Union NextGenerationEU/PRTR through the research project 2021/C005/0015007. Mr. Ronghao Pan is supported by the "Programa Investigo" grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2022a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

José Antonio García-Díaz, Francisco García-Sánchez, and Rafael Valencia-García. 2023a. Smart analysis of economics sentiment in spanish based on linguistic features and transformers. *IEEE Access*, 11:14211–14224.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2023b. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, 9(3):2893–2914.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

José Antonio García-Díaz, Pedro José Vivancos-Vicente, Angela Almela, and Rafael Valencia-García. 2022b. Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035–6044.

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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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. *CoRR*, abs/1907.11692.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human–differentiation analysis of scientific content generation. *arXiv*, 2301.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rexhep Shijaku and Ercan Canhasi. 2023. Chatgpt generated text detection. *Publisher: Unpublished*.

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.

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.