A Multi-Strategy Approach for AI-Generated Text Detection

Ali Zain
vin.alizain@gmail.com

Sareem Farooqui
sareemfarooqui10@gmail.com

Muhammad Rafi
muhammad.rafi@nu.edu.pk

National University of Computer and Emerging Sciences, FAST
Karachi, Pakistan

Abstract

This paper presents presents three distinct systems developed for the M-DAIGT shared task on detecting AI generated content in news articles and academic abstracts. The systems includes: (1) A fine-tuned RoBERTa-base classifier, (2) A classical TF-IDF + Support Vector Machine (SVM) classifier , and (3) An Innovative ensemble model named Candace, leveraging probabilistic features extracted from multiple Llama-3.2 models processed by a custom Transformer encoder.

The RoBERTa-based system emerged as the most performant, achieving near-perfect results on both development and test sets.

1 Introduction

The proliferation of sophisticated large language models (LLMs) has led to a surge in AI-generated text, making its detection a critical area of research (Jawahar et al., 2020). Identifying machine-generated content is crucial for maintaining information integrity, combating misinformation (Pan et al., 2023), and ensuring academic honesty. The M-DAIGT (Multi-domain DAIGT) shared task (Lamsiyah et al., 2025) aims to foster research in this domain by providing datasets for two distinct scenarios: news articles (Subtask 1) and academic abstracts (Subtask 2). Participants are tasked with building systems to classify given texts as either human-written or machine-generated.

In response to this challenge, our team developed and evaluated three different systems:

1. **RoBERTa-based Classifier:** A fine-tuned RoBERTa-base model, a widely successful approach for text classification tasks.

2. **TF-IDF + SVM Classifier:** A traditional machine learning pipeline combining Term Frequency-Inverse Document Frequency (TF-IDF) features with a Linear Support Vector Machine (SVM) (Joachims, 1998). This served as a strong baseline, particularly for Subtask 1.

3. **Llama-Feature Ensemble with Transformer Classifier (Candace):** An experimental system designed to capture nuanced signals from multiple LLMs. It extracts probabilistic features (Sarvazyan et al., 2024) (e.g., token log-probabilities, entropy) from a suite of Llama-3.2 models (Meta AI, 2024) and uses a custom Transformer Encoder-based model for final classification.

This paper details the architecture, data handling, implementation, and experimental results of these systems on the provided test datasets. Our RoBERTa-based approach yielded the most consistent and high-performing results on the development and test sets and was selected for our final submissions for both subtasks.

2 System Architectures

We developed three distinct systems, each employing a different strategy for AI-generated text detection.

2.1 System 1: RoBERTa-based Classifier

This system (Figure 1) fine-tunes a pre-trained RoBERTa-base model (Liu et al., 2019). The input text is tokenized, and the RoBERTa model processes these tokens. The final hidden state corresponding to the special ‘[CLS]‘ token is then passed through a linear classification layer to produce a binary prediction (human or machine).

2.2 System 2: TF-IDF + SVM Classifier

Our second system (Figure 2) follows a traditional machine learning pipeline. Textual input is first

Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text (M-DAIGT) associated with RANLP 2025, pages 15–19, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-108-0-003

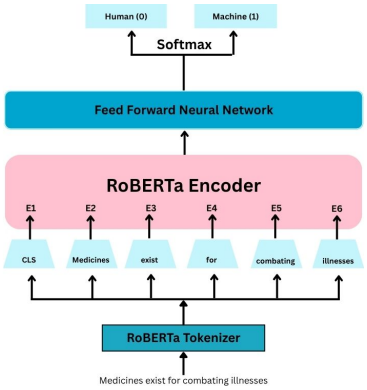

Figure 1: Architecture of System 1: RoBERTa-based Classifier.

converted into a numerical representation using TF-IDF vectorization, capturing n-grams. These TF-IDF features are then fed into a Linear Support Vector Machine (SVM) for classification.

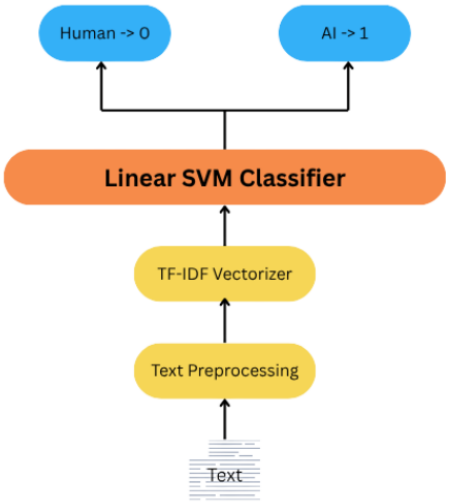

Figure 2: Architecture of System 2: TF-IDF + SVM Classifier.

2.3 System 3: Llama-Feature Ensemble with Transformer Classifier (Candace)

The third system (Figure 3), named Candace, is more experimental. It involves a two-stage process. First, probabilistic features (alpha, beta, gamma, as described in Section 4.3) are extracted from each token of the input text using multiple Llama-3.2 models. These feature vectors are concatenated. Second, this sequence of combined Llama-derived features is processed by a custom Transformer Encoder-based classification head, which then makes the final human/machine prediction.

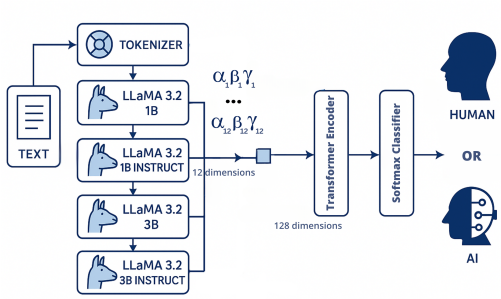

Figure 3: Architecture of System 3: Candace - Llama-Feature Ensemble with Transformer Classifier.

3 Data and Resources

The M-DAIGT shared task provided datasets for two subtasks:

- **Subtask 1 (News):** Comprised of ‘T1_train.csv‘ (10,000 samples), ‘T1_dev.csv‘ (2,000 samples), and ‘T1_test_unlabeled.csv‘ (2,000 samples).
- **Subtask 2 (Academic Abstracts):** Comprised of ‘T2_train.csv‘ (10,000 samples), ‘T2_dev.csv‘ (2,000 samples), and ‘T2_test_unlabeled.csv‘ (2,000 samples).

Each labeled dataset contained an ‘id‘, ‘text‘, and a ‘label‘ column, where labels were either ‘human‘ or ‘machine‘. For training, labels were mapped to 0 (human) and 1 (machine). Minimal preprocessing was applied for the RoBERTa and TF-IDF systems, primarily consisting of standard tokenization handled by the respective libraries. The Candace system’s feature extraction used raw text. External resources included:

- Pre-trained ‘roberta-base‘ model and tokenizer from Hugging Face Transformers (Wolf et al., 2020).
- Pre-trained Llama-3.2 models (Meta AI, 2024) (‘meta-llama/Llama-3.2-1B‘, ‘meta-llama/Llama-3.2-1B-Instruct‘, ‘meta-llama/Llama-3.2-3B‘, ‘meta-llama/Llama-3.2-3B-Instruct‘) and the ‘meta-llama/Llama-3.2-1B‘ tokenizer.

4 Methodology

4.1 System 1: RoBERTa-based Classifier

Model Architecture: We used the ‘RobertaModel‘ from Hugging Face Transformers, pre-

trained on ‘roberta-base‘. A linear classification layer was added on top of the pooled output (representation of the ‘[CLS]‘ token) from the RoBERTa model. The output layer predicts a score for the two classes (human vs. machine).

Input Representation: Texts were tokenized using ‘RobertaTokenizerFast‘ with a maximum sequence length of 512 tokens. Padding was applied to shorter sequences, and longer sequences were truncated.

Training: The model was fine-tuned for 4 epochs using the Adam optimizer with a learning rate of 1×10^{-5}. We used ‘CrossEntropyLoss‘ as the loss function. The batch size was set to 16. This setup was applied independently for both Subtask 1 and Subtask 2, using their respective training and development datasets.

4.2 System 2: TF-IDF + SVM Classifier

This system was developed primarily as a baseline for Subtask 1 (News).

Feature Extraction: We used `TfidfVectorizer` from scikit-learn to convert text into numerical features. We configured it to use n-grams of range (2, 3) and limited the maximum number of features to 5,000.

Classifier: A Linear Support Vector Machine (LinearSVC) was employed for classification. The hyperparameters were set as follows: C (regularization parameter) = 0.5, ‘class_weight=’balanced’‘ to handle potential class imbalance, ‘dual=False‘ (as n_samples were greater than n_features), and ‘max_iter=5000‘ to ensure convergence.

4.3 System 3: Llama-Feature Ensemble with Transformer Classifier (Candace)

This experimental system explores the utility of probabilistic features derived from multiple instruction-tuned and base Llama-3.2 models.

Feature Extraction: For each input text and for each of the four Llama models (‘meta-llama/Llama-3.2-1B‘, ‘meta-llama/Llama-3.2-1B-Instruct‘, ‘meta-llama/Llama-3.2-3B‘, ‘meta-llama/Llama-3.2-3B-Instruct‘), we extracted three features per token up to a maximum sequence length of 256:

- **Alpha (α):** The maximum log-probability assigned by the Llama model to any token at that position, given the preceding tokens.
- **Beta (β):** The entropy of the Llama model's predicted probability distribution over the vocabulary at that position.
- **Gamma (γ):** The log-probability assigned by the Llama model to the actual observed token at that position.

The Llama models were loaded with 8-bit quantization to manage memory. Features from all four Llama models were concatenated token-wise, resulting in 4 models $\times$ 3 features from each model $=$ 12 features per token.

Classifier Architecture (CandaceClassifier): The sequence of aggregated indicators was then processed by a custom classification architecture. This architecture begins with a projection of the indicator sequence into a higher-dimensional space. This transformed sequence is then passed through a Transformer Encoder block, designed to capture contextual relationships between the token-level indicators. The output of the Transformer Encoder is subsequently pooled across the sequence dimension, and a final linear layer produces the binary classification.

Training: The CandaceClassifier was trained for 10 epochs using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-4} and ‘CrossEntropyLoss‘. The batch size was 8. This architecture was trained separately for Subtask 1 and Subtask 2.

5 Experiments and Results

All systems were trained and evaluated on the M-DAIGT development sets for their respective subtasks. The primary evaluation metrics were Accuracy and F1-score.

RoBERTa-based System (System 1): For Subtask 1 (News), our fine-tuned RoBERTa model achieved an accuracy of 99.95% and an F1-score of 99.95% on the development set (best at epoch 4). For Subtask 2 (Academic Abstracts), the RoBERTa model achieved 100.00% accuracy and 100.00% F1-score on the development set (stable from epoch 1 onwards). Given its strong and consistent performance, this system was chosen for our official submissions for both subtasks.

TF-IDF + SVM System (System 2): This system was evaluated on Subtask 1 and Subtask 2. On

System	Subtask 1 (News) - Test Set				Subtask 2 (Academic) - Test Set			
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
RoBERTa-base (System 1)	**99.99%**	**99.99%**	**99.80%**	**100.0%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
TF-IDF + SVM (System 2)	97.90%	97.91%	97.52%	98.30%	99.85%	99.85%	100.0%	99.70%
Candace (System 3)	99.75%	99.75%	99.60%	99.90%	99.95%†	99.95%	100.00%	99.90%

Table 1: Test set performance, RoBERTa base model with Fast tokenizer outperforming all models

its internal training data (as dev metrics were not explicitly separated in its notebook), it achieved an accuracy of 99.81% and an F1-score of 0.9981. While competitive, it was slightly outperformed by the RoBERTa model on the development set.

Candace System (System 3): For Subtask 1 (News), the Candace system achieved a development accuracy of 99.80% (best at epoch 6). The same architectural design and training procedure were applied to Subtask 2, and similar development accuracy (99.80%) was observed during its separate training run. While promising, this system is more computationally intensive due to the multi-LLM feature extraction step. The RoBERTa system offered slightly better or comparable performance with significantly less overhead for these specific datasets.

6 Discussion

Our experiments highlight the continued effectiveness of fine-tuned transformer models like RoBERTa for text classification tasks, achieving near-perfect scores on the development sets for both news and academic abstract domains. The RoBERTa model's ability to capture subtle linguistic cues makes it highly suitable for distinguishing between human and AI-generated text.

The TF-IDF + SVM approach, while simpler, provided a very strong baseline for Subtask 1, underscoring the utility of traditional methods, especially when coupled with robust feature engineering like n-grams.

The Candace system, which extracts features from multiple Llama-3.2 models, also showed excellent performance. This approach is interesting as it attempts to distill knowledge from several powerful LLMs into a smaller, specialized classifier. However, the feature extraction process is computationally expensive. For the M-DAIGT datasets, the gains over a well-tuned RoBERTa model were not substantial enough to justify the additional complexity and computational cost as the primary submission.

Runtime for RoBERTa inference is efficient, while Candace inference is slower due to the initial pass through multiple Llama models.

7 Conclusion

We presented three distinct systems for detecting AI-generated text in news articles and academic abstracts. Our fine-tuned RoBERTa-base model demonstrated exceptional performance on the development and test sets for both subtasks, achieving near-perfect accuracy and F1-scores, and was selected as our primary submission. The TF-IDF+SVM system served as a strong baseline, and the experimental Candace system, leveraging features from multiple Llama models, also showed high efficacy. Future work could involve ensembling these diverse models, exploring more sophisticated feature fusion techniques for the Candace system, and investigating the robustness of these models against adversarial attacks or text generated by newer, more advanced language models.

Acknowledgments

We thank the organizers of the M-DAIGT shared task for providing the dataset and the evaluation platform. BEsides that, the research is also supported by the provisional award under the National Research Program for Universities (NRPU), Higher Education Commission (HEC) Pakistan, with the title "NRPU: Automatic Multi-Model Classification of Religious Hate Content from Social Media" (Reference Research Project No. 16153).

References

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant

features. In *European conference on machine learning*, pages 137–142. Springer.

Salima Lamsiyah, Saad Ezzini, Abdelkader Elmahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El Amrany, Salmane Chafik, and Hicham Hammouchi. 2025. Shared task on multi-domain detection of ai-generated text (m-daigt). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

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. In *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. ArXiv:1711.05101.

Meta AI. 2024. Llama 3.2 Model Family. https://www.llama.com/models/llama-3/.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Arsen M. Sarvazyan, Jorge Á. González, and Marc Franco-Salvador. 2024. Genaios at SemEval-2024 task 8: Detecting machine-generated text by mixing language model probabilistic features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.