

Generative AI Text Classification using Ensemble LLM Approaches

Harika Abburi^{1,*}, Michael Suesserman², Nirmala Pudota¹, Balaji Veeramani², Edward Bowen² and Sanmitra Bhattacharya²

¹Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited, India

²Deloitte & Touche LLP, USA

Abstract

Large Language Models (LLMs) have shown impressive performance across a variety of Artificial Intelligence (AI) and natural language processing tasks, such as content creation, report generation, etc. However, unregulated malign application of these models can create undesirable consequences such as generation of fake news, plagiarism, etc. As a result, accurate detection of AI-generated language can be crucial in responsible usage of LLMs. In this work, we explore 1) whether a certain body of text is AI generated or written by human, and 2) attribution of a specific language model in generating a body of text. Texts in both English and Spanish are considered. The datasets used in this study are provided as part of the Automated Text Identification (AuTextTification) shared task. For each of the research objectives stated above, we propose an ensemble neural model that generates probabilities from different pre-trained LLMs which are used as features to a Traditional Machine Learning (TML) classifier following it. For the first task of distinguishing between AI and human generated text, our model ranked in fifth and thirteenth place (with macro $F1$ scores of 0.733 and 0.649) for English and Spanish texts, respectively. For the second task on model attribution, our model ranked in first place with macro $F1$ scores of 0.625 and 0.653 for English and Spanish texts, respectively.

Keywords

generative AI, text classification, large language models, ensemble

1. Introduction

Rapid advances in the capabilities of LLMs, and their ease of use in generating sophisticated and coherent content is leading to the production of AI-generated content at scale. Some of the applications of LLMs are in AI-assisted writing [1], medical question answering [2, 3] and a wide range of tasks in the financial services industry [4] and legal domain [5]. Foundational models such as OpenAI's GPT-3 [6], Meta's OPT [7], and Big Science's BLOOM [8] can generate such sophisticated content with basic text prompts that it is often challenging to manually discern between human and AI-generated text. While these models demonstrate the ability to understand the context and generate coherent human-like responses, they do not have a true


IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

✉ abharika@deloitte.com (H. Abburi); msuesserman@deloitte.com (M. Suesserman); npudota@deloitte.com (N. Pudota); bveeramani@deloitte.com (B. Veeramani); edbowen@deloitte.com (E. Bowen); sanmbhattacharya@deloitte.com (S. Bhattacharya)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

understanding of what they are producing [9, 10]. This could potentially lead to adverse consequences when used in downstream applications. For example, consider an application of a LLM used for summarizing a medicinal drug data-sheet that inadvertently produces wrong dosage information. Generating plausible but false content (referred to as *hallucination* [11, 12]), may inadvertently help propagate misinformation, false narratives, fake news, and spam. Given the pace of LLM adoption by the general public, and the rate of dissemination of information across the globe, widespread misinformation propagation is an imminent risk that both individuals and organizations will have to deal with in the near future [13, 14]. An understanding of the source of the authored content – whether by an AI system or a human – would allow one to appropriately use the content in downstream applications with suitable oversight. In the case of AI-generated content, the knowledge of the source LLM would allow one to watch out for potential known biases and limitations associated with that model. Given the risks associated with unchecked and unregulated adoption of generative AI models, it is imperative to develop approaches to detect AI-generated content, and identify the source of the AI-generated content.

Motivated by these challenges, automatic detection of AI-generated text has become an active area of research. Recent work such as DetectGPT [15] generates minor perturbations of a passage using a generic pre-trained Text-to-Text Transfer Transformer (T5) model, and then compares the log probability of the original sample with each perturbed sample to determine if it is AI generated. Another approach [16] developed a model for a generative multi-class AI detection challenge involving Russian language text using Decoding- enhanced BERT with disentangled attention (DeBERTa) as a pre-trained language model for classification and an ensemble approach is proposed by [17] for binary classification. Mitrovic *et al.* [18] developed a fine-tuned transformer-based approach to distinguish between human and ChatGPT generated text, with addition of SHapely Additive exPlanations (SHAP) values for model explainability. Statistical detection methods have also been applied for detection of generative AI text, such as the Giant Language model Test Room (GLTR) approach developed by researchers [19]. Research involving repeated higher-order n-grams found that they occur more often in AI generated text as compared to human generated text [20]. Using an ensemble of classifiers, higher-order n-grams can be used to help distinguish between human and AI generated text.

To boost this area of research further, the Automated Text Identification (AuTextification) [21] shared task, part of IberLEF 2023 [22], put forth two tasks, each covering both English and Spanish language texts: 1. *Human or generated*: determine whether a given input text has been automatically generated or not – a binary classification task with two classes ‘human’ or ‘AI’ 2. *Model attribution*: for automatically generated text, determine which one of six different text generation AI models was the source – a multi-class classification task with six classes, where each class represents a text generation model. For each of these tasks, we propose an ensemble classifier, where the probabilities generated from various state-of-the-art LLMs are used as input feature vectors to TML models to produce final predictions. Our experiments show multiple instances of the proposed framework outperforming several baselines using established evaluation metrics.

2. Dataset

The data for all the subtasks and languages, is provided by the AuTextification shared task organizers [21]. For Task 1, *Human or generated*, the data consists of texts from five domains. During the training phase of the shared task the domains are not disclosed to the participants. For Task 1, for both English and Spanish, data from three domains is provided as training data, and data from two new domains is provided in the test set. For Task 2, *Model attribution*, also the data comes from five different domains, but the same domains are in train and test splits. This data is evenly distributed into six class (A, B, C, D, E, and F), where each class represents a text generation model. An interesting point to note here is that the six text generation models are of increasing number of neural parameters, ranging from 2B to 175B. The motivation here is to emulate realistic AI text detection approaches which should be versatile enough to detect a diverse set of text generation models and writing styles. More details about the data can be found in the AuTextification overview paper [21].

3. Proposed Ensemble Approach

In this section, we describe our approaches for detection and classification of generative AI text.

3.1. Models

We explored various state-of-the-art large language models [23] such as Bidirectional Encoder Representations from Transformers (BERT), DeBERTa, Robustly optimized BERT approach (RoBERTa), and cross-lingual language model RoBERTa (XLM-RoBERTa) along with their variants. Since the datasets are different for each task and language, and the same set of the models will not fit across them, we fine-tuned different models for different subtasks and languages and pick the best models based on validation data. Table 1 lists the different models that we explored for the different subtasks: Task 1 in English (Binary-English), Task 1 in Spanish (Binary-Spanish), Task 2 in English (Multiclass-English) and Task 2 in Spanish (Multiclass-Spanish). We also explored different TML models such as Linear SVC [24], Error-Correcting Output Codes (ECOC) [25], OneVsRest [26] and Voting classifier which includes Logistic Regression (LR), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector machines (SVC) [27] in our approach and the best model for each subtask is shown in results section.

3.2. Proposed Ensemble Approach

Each input text is passed through variants of the pre-trained large language models such as DeBERTa (D), XLM-RoBERTa (X), RoBERTa (R), BERT (B), etc. During the model training phase, these models are fine-tuned on the training data. For inference and testing, each of these models can independently generate classification probabilities (P), namely P^D , P^X , P^R , P^B , etc. These probabilities are concatenated (P^C) or averaged (P^A), and the output is passed as a feature vector to train TML models to produce final predictions.

Table 1

Models explored for different tasks

Task	Large language models
Binary-English	deberta-large [28], xlm-r-100langs-bert-base-nli-stsb-mean-tokens [29], roberta-base-openai-detector [30], xlm-roberta-large-xnli-anli, roberta-large
Binary-Spanish	bertin-roberta-base-spanish [31], MarIA [32], sentence_similarity_spanish_es, xlm-roberta-large-xnli-anli, xlm-roberta-large-finetuned-conll02-spanish [33]
Multiclass-English	xlm-roberta-large-finetuned-conll03-english, scibert_scivocab_cased [34], deberta-base, roberta-large [35], longformer-base-4096 [36], bert-large-uncased-whole-word-masking-finetuned-squad [37]
Multiclass-Spanish	xlm-roberta-large-finetuned-conll03-english, MarIA, sentence_similarity_spanish_es, bert-base-multilingual-cased-finetuned-conll03-spanish, roberta-large

4. Experiments

This section provides the experimental evaluation of the proposed methods. For both the tasks we report results on accuracy (Acc), macro F1 score (F_{macro}), precision ($Prec$) and recall (Rec).

4.1. Implementation Details

We set aside 20% from the training data for validation. For the testing phase, the validation set is merged with the training set. All the results are reported on testing data. The hyper-parameters used for model fine-tuning are shown in Table 2.

Table 2

Hyper-parameters for all the subtasks

Parameter	value
Batch size	128
Learning rate	0.00003
Maximum sequence length	128
Epochs	10 for task1 and 20 for task2

4.2. Results

For each task and language, we submitted three runs to the leaderboard (team name *Drocks*). These runs correspond to the most promising approaches on the validation data. In this paper, we show only the top run results for each of the tasks. The complete leaderboard is available at <https://sites.google.com/view/autextification/results> [21]. The results on test data for Binary-English and Binary-Spanish are shown in Tables 3 and 4, respectively. With the concatenated feature vector (P^C) as an input, a voting classifier results in F_{macro} score of 73.3 on Binary-English

Table 3
Results for the Binary-English task

Model	Acc	F_{macro}	Prec	Rec
deberta-large	0.620	0.546	0.783	0.610
xlm-r-100langs-bert-base-nli-stsb-mean-tokens	0.647	0.592	0.782	0.639
roberta-base-openai-detector	0.679	0.636	0.805	0.671
xlm-roberta-large-xnli-anli	0.618	0.543	0.782	0.608
roberta-large	0.623	0.551	0.784	0.613
Ensemble with Voting classifier (P^C as a input feature)	0.751	0.733	0.826	0.745

Table 4
Results for the Binary-Spanish task

Model	Acc	F_{macro}	Prec	Rec
bertin-roberta-base-spanish	0.698	0.633	0.798	0.661
MarIA	0.690	0.629	0.791	0.652
sentence_similarity_spanish_es	0.651	0.560	0.786	0.607
xlm-roberta-large-xnli-anli	0.633	0.526	0.788	0.587
xlm-roberta-large-finetuned-conll02-spanish	0.637	0.533	0.787	0.591
Ensemble with OneVsRest classifier (P^C as a input feature)	0.704	0.649	0.805	0.667

Table 5
Results for the Multi-English task

Model	Acc	F_{macro}	Prec	Rec
xlm-roberta-large-finetuned-conll03-english	0.598	0.593	0.618	0.594
scibert_scivocab_cased	0.578	0.576	0.590	0.575
deberta-base	0.564	0.558	0.602	0.558
roberta-large	0.581	0.568	0.611	0.574
longformer-base-4096	0.586	0.582	0.600	0.582
bert-large-uncased-whole-word-masking-finetuned-squad	0.581	0.581	0.597	0.579
Ensemble with ECOC classifier (P^C as a input feature)	0.624	0.625	0.649	0.621

data, whereas a onestsrest classifier outperforms other methods with F_{macro} score of 64.9 on Binary-Spanish data.

Tables 5 and 6 show the results on test data for Multiclass-English and Multiclass-Spanish, respectively. For the Multiclass-English data, an ECOC classifier on top of the concatenated feature vector (P^C) outperforms the other approaches with F_{macro} score of 62.5. On the other hand, a linear SVC classifier with averaged feature vector (P^A) as an input outperforms the other approaches with F_{macro} score of 65.4 for the Multiclass-Spanish data.

Table 6
Results for the Multi-Spanish task

Model	Acc	F_{macro}	Prec	Rec
xlm-roberta-large-finetuned-conll03-english	0.632	0.629	0.661	0.628
MarIA	0.614	0.615	0.630	0.612
sentence_similarity_spanish_es	0.615	0.612	0.640	0.613
bert-base-multilingual-cased-finetuned-conll03-spanish	0.593	0.594	0.599	0.593
roberta-large	0.584	0.584	0.595	0.584
Ensemble with Linear SVC classifier (P^A as a input feature)	0.653	0.654	0.679	0.650

5. Conclusion

In this paper, we described our submission to the AuTextification shared task which consists of two tasks on the classification of generative AI content. In our experiments, we found that our proposed ensemble LLM approach is a promising strategy, as our model ranked fifth with a macro $F1$ score of 73.3% for English and thirteenth with 64.9% macro $F1$ score for Spanish in the *Human or generated* binary classification task. For the *Model attribution* multiclass classification task, our model ranked in the first place for both English and Spanish, with macro $F1$ scores of 62.5% and 65.3%, respectively. While our approach shows promising results for the *Model attribution* task, further exploration is needed to enhance and tune our models for *Human or generated* binary classification task.

References

- [1] M. Hutson, Robo-writers: the rise and risks of language-generating ai, *Nature* 591 (2021) 22–25.
- [2] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, *PLoS digital health* 2 (2023) e0000198.
- [3] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, J. Mostafa, A systematic review of automatic text summarization for biomedical literature and ehers, *Journal of the American Medical Informatics Association* 28 (2021) 2287–2297.
- [4] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, Bloomberggpt: A large language model for finance, *arXiv preprint arXiv:2303.17564* (2023).
- [5] Z. Sun, A short survey of viewing large language models in legal aspect, *arXiv preprint arXiv:2303.09136* (2023).
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.

- [7] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, arXiv preprint arXiv:2205.01068 (2022).
- [8] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).
- [9] H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, J. W. Gichoya, Ethics of large language models in medicine and medical research, *The Lancet Digital Health* (2023).
- [10] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, arXiv preprint arXiv:2305.04388 (2023).
- [11] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al., A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, arXiv preprint arXiv:2302.04023 (2023).
- [12] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Computing Surveys* 55 (2023) 1–38.
- [13] H. Ali, J. Qadir, Z. Shah, Chatgpt and large language models (llms) in healthcare: Opportunities and risks (2023).
- [14] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359 (2021).
- [15] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, arXiv preprint arXiv:2301.11305 (2023).
- [16] B. Li, Y. Weng, Q. Song, H. Deng, Artificial text detection with multiple training strategies, arXiv preprint arXiv:2212.05194 (2022).
- [17] N. Maloyan, B. Nutfullin, E. Ilyushin, Dialog-22 ruatd generated text detection, arXiv preprint arXiv:2206.08029 (2022).
- [18] S. Mitrović, D. Androletti, O. Ayoub, Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text, arXiv preprint arXiv:2301.13852 (2023).
- [19] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text, arXiv preprint arXiv:1906.04043 (2019).
- [20] M. Gallé, J. Rozen, G. Kruszewski, H. Elshahar, Unsupervised and distributional detection of machine-generated text, arXiv preprint arXiv:2111.02878 (2021).
- [21] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains, in: *Procesamiento del Lenguaje Natural*, Jaén, Spain, 2023.
- [22] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, *Procesamiento del Lenguaje Natural* 71 (2023).
- [23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 conference on empirical methods in natural language processing*:

system demonstrations, 2020, pp. 38–45.

- [24] S. Sulaiman, R. A. Wahid, A. H. Ariffin, C. Z. Zulkifli, Question classification based on cognitive levels using linear svc, *Test Eng Manag* 83 (2020) 6463–6470.
- [25] K.-H. Liu, J. Gao, Y. Xu, K.-J. Feng, X.-N. Ye, S.-T. Liong, L.-Y. Chen, A novel soft-coded error-correcting output codes algorithm, *Pattern Recognition* 134 (2023) 109122.
- [26] J.-H. Hong, S.-B. Cho, A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification, *Neurocomputing* 71 (2008) 3275–3281.
- [27] A. Mahabub, A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers, *SN Applied Sciences* 2 (2020) 525.
- [28] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: *International Conference on Learning Representations*, 2021. URL: <https://openreview.net/forum?id=XPZlaotutsD>.
- [29] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. URL: <http://arxiv.org/abs/1908.10084>.
- [30] I. Solaiman, M. Brundage, J. Clark, A. Askill, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al., Release strategies and the social impacts of language models, *arXiv preprint arXiv:1908.09203* (2019).
- [31] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, *Procesamiento del Lenguaje Natural* 68 (2022) 13–23. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403>.
- [32] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* 68 (2022). URL: <https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley>. doi:10.26342/2022-68-3.
- [33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *arXiv preprint arXiv:1911.02116* (2019).
- [34] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: *EMNLP*, Association for Computational Linguistics, 2019. URL: <https://www.aclweb.org/anthology/D19-1371>.
- [35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, *CoRR abs/1907.11692* (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [36] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, *arXiv:2004.05150* (2020).
- [37] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *CoRR abs/1810.04805* (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.