An Ensemble-Based Approach for Generative Language Model Attribution

Harika Abburi[1(✉)], Michael Suesserman[2], Nirmala Pudota[1], Balaji Veeramani[2], Edward Bowen[2], and Sanmitra Bhattacharya[2]

[1] Deloitte and Touche Assurance and Enterprise Risk Services India Private Limited, Hyderabad, India
{abharika,npudota}@deloitte.com
[2] Deloitte and Touche LLP, New York, USA
{msuesserman,bveeramani,edbowen,sanmbhattacharya}@deloitte.com

Abstract. Recently, Large Language Models (LLMs) have gained considerable attention due to their incredible ability to automatically generate texts that closely resemble human-written text. They have become invaluable tools in handling various text-based tasks such as content creation and report generation. Nevertheless, the proliferation of these tools can create undesirable consequences such as generation of false information and plagiarism. A variety of LLMs have been operationalized in the last few years whose abilities are heavily influenced by the quality of their training corpus, model architecture, pre-training tasks, and fine-tuning processes. Our ability to attribute the generated text to a specific LLM will not only help us understand differences in the LLMs' output characteristics, but also effectively distinguish machine-generated text from human-generated text. In this paper, we study whether a machine learning model can be effectively trained to attribute text to the underlying LLM that generated it. We propose an ensemble neural model that generates probabilities from multiple pre-trained LLMs, which are then used as features for a traditional machine learning classifier. The proposed approach is tested on Automated Text Identification (AuTexTification) datasets in English and Spanish languages. We find that our models outperform various baselines, achieving macro F_{macro} scores of 0.63 and 0.65 for English and Spanish texts, respectively.

Keywords: Generative AI · Model Attribution · Large language models · Ensemble

1 Introduction

Recent advancements in machine learning and natural language processing research have paved the way for the development of sophisticated LLMs. The widespread availability and the ease with which they can generate coherent content are contributing to the production of massive volumes of automatically generated online content. LLMs have demonstrated remarkable performance in

F. Zhang et al. (Eds.): WISE 2023, LNCS 14306, pp. 699–709, 2023.
https://doi.org/10.1007/978-981-99-7254-8_54

producing human-like language, showcasing their potential use across a wide range of applications, such as domain specific tasks in legal [20] and financial services [23]. Foundation models such as OpenAI's GPT-3 [1] and Big Science's Bloom [19] are publicly available, and can generate highly sophisticated content with basic text prompts. This often presents a challenge to discern between human and LLM-generated text.

While LLMs demonstrate the ability to understand the context and generate coherent human-like responses, they do not have a true understanding of what they are producing [12]. This could potentially lead to adverse consequences when used in downstream applications. Generating plausible but false content (*hallucination* [10]), may inadvertently help propagate misinformation, fake news, and spam [9].

There is a considerable body of research available on detecting text generated by artificial intelligence (AI) systems [9,21]. However, the identification of a specific LLM responsible for generating such text is a relatively new area of research. We argue that attributing the generated text to a specific LLM is a vital research area, as the knowledge of the source LLM would enable one to be vigilant regarding potential known biases and limitations associated with that model and use the content appropriately in downstream applications with suitable oversight [21].

In this study, we focus on identifying the source of the AI-generated text (referred to as model attribution hereafter) in two different languages, English and Spanish. More specifically, given a piece of text, the goal is to determine which specific LLM generated the text. To address this problem statement, we propose an ensemble classifier, where the probabilities generated from various state-of-the-art LLMs are used as input feature vectors to traditional machine learning classification models to produce the final predictions. Our experiments show multiple instances of the proposed framework outperform several baselines using well-established evaluation metrics.

2 Related Work

The majority of research in this area is focused on differentiating between text authored by humans and text generated by AI [3,17].

The use of neural networks leveraging complex linguistic features and their derivatives is most prevalent in detecting AI-generated text. 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. Deng *et al.* [4] build upon the DetectGTP model by incorporating a Bayesian surrogate model to select text samples more efficiently, which achieves similar performance as DetectGTP using half the number of samples. Mitrovic *et al.* [16] developed a fine-tuned Transformer-based approach to distinguish between human and ChatGTP generated text, with the addition of SHapely Additive exPlanations (SHAP) values for model explainability. This approach provides insight into the reasoning behind the model's predictions. Statistical methods have also been applied for detection of AI-generated text, such as the Giant Language model Test Room (GLTR) approach [6].

The increasing sophistication of generative AI models coupled with adversarial attacks make detection of AI-generated text especially challenging. Two forms of attacks that create additional complications are paraphrasing attacks and adversarial human spoofing [17]. Automatically generated text may also show factual, grammatical, or coherence artifacts [14] along with statistical abnormalities that impact the distributions of automatic and human texts [8]. The importance of detecting AI-generated text and the corresponding challenges will foster further research on this topic.

In addition to distinguishing between human and AI-generated text, identifying a specific LLM that generates the artificial text is becoming increasingly important. Uchendu *et al.* [21] explored the Robustly optimized BERT approach (RoBERTa) model to classify AI-generated text into eight different classes. Li *et al.* [11] developed a model for AI-generated multi-class text classification on Russian language using Decoding-enhanced BERT with disentangled attention (DeBERTa) as a pre-trained language model for category classification. These prior works focused on model attribution for only a single language, such as English or Russian. In contrast to the aforementioned research, and to the extent of our knowledge, our approach to model attribution is the first one to be applied across multiple languages, demonstrating the robustness of our approach across attributable LLMs, languages, and domains.

3 AuTexTification Dataset

The dataset used in the study comes from the Iberian Languages Evaluation Forum (IberCLEF)-AuTexTification shared task [18]. The data consists of texts from five domains, where three domains (legal, wiki, and tweets) are used for training, and two different domains are used for testing (reviews and news). It contains machine generated text from six text generation models, labeled as bloom-1b7 (A), bloom-3b (B), bloom-7b1 (C), babbage (D), curie (E), and text-davinci-003 (F) for two different languages, English and Spanish. The LLMs used to generate the text are of increasing number of neural parameters, ranging from 2B to 175B. The motivation here is to emulate realistic AI text detection approaches that should be versatile enough to detect a diverse set of text generation models and writing styles. The number of samples in each class for both languages is shown in Table 1. To showcase the complexity of the problem, we also present samples for each category from both the English and Spanish datasets in Tables 2 and 3.

4 Proposed Ensemble Approach

In this Section, we detail our approach for conducting the generative language model attribution. We first provide a description of the LLMs and machine learning models that we explored for model attribution. Next, we discuss the proposed

Table 1. Label distribution across the languages for model attribution task. Train and test splits for each language are also shown.

Category	Multiclass-English		Multiclass-Spanish	
	Train	*Test*	*Train*	*Test*
bloom-1b7 (A)	3562	887	3422	870
bloom-3b (B)	3648	875	3514	867
bloom-7b1 (C)	3687	952	3575	878
babbage (D)	3870	924	3788	946
curie (E)	3822	979	3770	1004
text-davinci-003 (F)	3827	988	3866	917

Table 2. Samples of English AI-generated text, with corresponding source models (labeled A-F).

Text	Label
The best songs are those that I can sing along with, and they're all there!	B
Summer Vacation time. That means we will have to get the kids in a school setting. We can look forward to all of that	C
Thanks @arohan and @MoneyEnergy I have heard this argument many times before. Im not going to get into it here, but suffice	E

Table 3. Samples of Spanish AI-generated text, with corresponding source models (labeled A-F).

Text	Label
¿En qué se parecen las ofertas de trabajo y los puestos laborales que están disponibles para el trabajador colombiano en este momento?. Las opciones	A
No se trata de ser una revolución, pero de que la gente sienta que ha visto un cambio y que quiere seguir segu	D
Los padres pueden negar la solicitud del niño y ofrecer alternativas saludables, como frutas, verduras o una bebida sin cafeína. Además, los padres pueden explicar por qué es importante que los niños consuman alimentos saludables y cuáles son los efectos negativos de comer alimentos poco saludables	F

Table 4. Models explored for English and Spanish datasets

Task	Large language models
English	xlm-roberta-large-finetuned-conll03-english, allenai/scibert_scivocab_cased, microsoft/deberta-base, roberta-large, allenai/longformer-base-4096, bert-large-uncased-whole-word-masking-finetuned-squad
Spanish	xlm-roberta-large-finetuned-conll03-english, PlanTL-GOB-ES/roberta-large-bne, hiiamsid/sentence_similarity_spanish_es, dbmdz/bert-base-multilingual-cased-finetuned-conll03-spanish, roberta-large

ensemble neural architecture, where we fine-tuned the LLMs and then passed their predictions to various traditional machine learning models to perform the ensemble operation.

4.1 Models

LLMs: We explored various state-of-the-art LLMs [22], such as Bidirectional Encoder Representations from Transformers (BERT), DeBERTa, RoBERTa, and cross-lingual language model RoBERTa (XLM-RoBERTa) along with their variants. Since the datasets are different for each language, and the same set of models will not fit across them, we fine-tuned different models for different languages. We investigated more than 15 distinct models for each language and selected the ones presented in this paper based on their performance on the validation data. This selection was made to ensure model diversity, which aids in generalisation and improved comprehension of context and semantics. Table 4 lists the different models that we selected for the two languages under consideration. We briefly describe each of the LLMs below.

- **microsoft/deberta-base** [7] is a transformer model which improves the BERT and RoBERTa models using disentangled attention and enhanced mask decoder.
- **xlm-roberta-large-finetuned-conll03-english** is XLM-RoBERTa based model [2] which is a large multi-lingual language model trained on 2.5TB of filtered Common Crawl data. The conll03-english model is fine-tuned on the XLM-RoBERTa model with conll2003 dataset in English.
- **roberta-large, PlanTL-GOB-ES/roberta-large-bne** are RoBERTa based models [13] which are pre-trained on a large corpus of English data in a self-supervised fashion using a Masked Language Modeling (MLM) objective. The roberta-large-bne model has been pre-trained using the largest Spanish corpus with a total of 570GB of text compiled from the web crawlings.
- **dbmdz/bert-base-multilingual-cased-finetuned-conll03-spanish, hiiamsid/sentence_similarity_spanish_es, allenai/scibert_scivocab_cased, bert-large-uncased-whole-word-masking-finetuned-squad, and allenai/longformer-base-4096** are BERT-based models [5]. The bert-base-multilingual model is pre-trained on 104 languages with the largest Wikipedia data using a MLM objective and further pre-trained on the CoNLL-2002 dataset in Spanish. The *sentence similarity* Spanish model is a sentence-transformer model where the base model is BETO which is trained on a large Spanish corpus. The scibert model is trained on papers taken from Semantic Scholars. The BERT-large SQuAD model is slightly different from other BERT models since it is trained with a whole word masking technique and further fine-tuned on the Stanford Question Answering Dataset (SQuAD). The Long-Document Transformer (Longformer) model is a BERT-like model stemmed from the RoBERTa checkpoint and pre-trained for MLM on long documents which supports sequences of lengths up to 4,096.

Machine Learning (ML) Models: We explored various traditional machine learning and ensembling models such as Bagging , Voting, OneVsRest, Error-Correcting Output Codes (ECOC), and LinearSVC [24].

4.2 Proposed Ensemble Neural Architecture

As shown in Fig. 1, an input text is passed through variants of the pre-trained LLMs such as, DeBERTa (D), XLM-RoBERTa (X), RoBERTa (R), and BERT (B). During the model training phase, these models are fine-tuned on the training data. For inference and testing, each of these models independently generate classification probabilities (P), namely P^D, P^X, P^R, P^B, etc. In order to maximize the contribution of each model, each of these probabilities are concatenated (P^C) or averaged (P^A), and this output is passed as a feature vector to train various traditional ML models to produce final predictions.

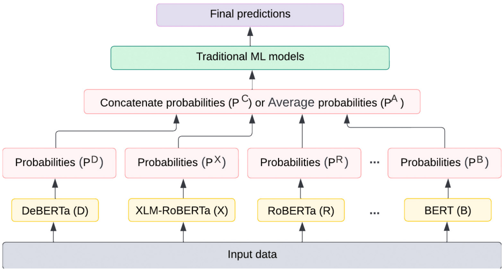

Fig. 1. Proposed ensemble neural architecture

5 Experiments

In this section, we discuss the evaluation of the proposed methods. We report model performance using well-established metrics such as accuracy (Acc), macro F1 score (F_{macro}), precision ($Prec$) and recall (Rec).

5.1 Baselines

We establish Linear Support Vector Classification (SVC), Logistic Regression (LR), and Random Forests (RF) as baselines, where each baseline model takes two distinct feature sets – word n-grams and character n-grams. We also explored other baselines like the Symanto Brain Few-shot and Zero-shot without label verbalization approaches[1], but due to their relatively low performance compared to the approaches presented in Table 5, we do not report those results.

[1] https://www.symanto.com/nlp-tools/symanto-brain/.

5.2 Implementation Details

During model training we set aside 20% from the training data for validation. However, for the held-out testing phase, the validation set is merged with the training set. The following hyper-parameters are used for model fine-tuning: batch size - 128, learning rate - $3e^{-}5$, max sequence length - 128, and number of epochs is set to 20. We also used a sliding window to prevent the truncation of longer sequences, allowing the model to handle longer sentences.

Table 5. Baseline results of model attribution for both English and Spanish.

Classifier	Features	English				Spanish			
		Acc	F_{macro}	*Prec*	*Rec*	*Acc*	F_{macro}	*Prec*	*Rec*
Linear SVC	word n-grams	0.360	0.355	0.354	0.357	0.464	0.459	0.457	0.463
	character n-grams	0.439	0.428	0.425	0.438	**0.505**	0.495	0.493	**0.505**
LR	word n-grams	0.374	0.368	0.366	0.371	0.482	0.475	0.473	0.481
	character n-grams	**0.451**	**0.440**	**0.438**	**0.450**	**0.505**	**0.496**	**0.495**	**0.505**
RF	word n-grams	0.339	0.330	0.330	0.337	0.425	0.407	0.409	0.425
	character n-grams	0.414	0.400	0.399	0.413	0.437	0.423	0.423	0.436

Table 6. Results of model attribution on the English dataset

Model	*Acc*	F_{macro}	*Prec*	*Rec*
xlm-roberta-large-finetuned-conll03-english	0.598	0.593	0.618	0.594
allenai/scibert_scivocab_cased	0.578	0.576	0.590	0.575
microsoft/deberta-base	0.564	0.558	0.602	0.558
roberta-large	0.581	0.568	0.611	0.574
allenai/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 P^C as an input feature				
Bagging	0.597	0.599	0.614	0.595
voting	0.607	0.603	0.650	0.603
OneVsRest	0.625	0.626	**0.651**	0.622
output code	0.624	0.625	0.649	0.621
Linear SVC	**0.629**	**0.630**	0.637	**0.626**

Table 7. Results of model attribution on the Spanish dataset

Model	*Acc*	F_{macro}	*Prec*	*Rec*
xlm-roberta-large-finetuned-conll03-english	0.632	0.629	0.661	0.628
PlanTL-GOB-ES/roberta-large-bne	0.614	0.615	0.630	0.612
hiiamsid/sentence_similarity_spanish_es	0.615	0.612	0.640	0.613
dbmdz/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 P^C **as an input feature**				
Bagging	0.616	0.615	0.637	0.613
voting	0.631	0.630	**0.691**	0.627
OneVsRest	0.648	0.648	0.677	0.645
output code	0.646	0.647	0.677	0.643
Linear SVC	**0.655**	**0.656**	0.669	**0.652**

5.3 Results

Table 5 shows results produced using three traditional ML methods (Linear SVC, LR, and RF) across two different feature sets (word n-grams and character n-grams) for both languages. LR with character n-grams outperforms other approaches on the macro $F1$ performance metric for both languages.

Tables 6 and 7 provide results on English and Spanish datasets respectively, with different variants of the proposed architecture. The first block in the table shows the results for individual LLMs. The second and third blocks show the ensemble results with P^C and P^A respectively, as input feature vector to several machine learning models.

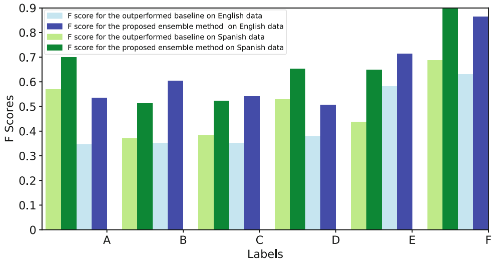

Fig. 2. Class-wise F-scores for the outperformed baseline (*LR with character n-grams*) and proposed ensemble method (*Linear SVC*) on English dataset

The results on the English test data are shown in Table 6. Out of all the combinations, Linear SVC with concatenated feature vector (P^C) as an input, outperforms other approaches for a majority of the evaluation metrics with an F_{macro} score of 0.63. Table 7 shows the results on the Spanish test dataset where the concatenated feature vector (P^C) is passed as an input to the Linear SVC classifier outperforms the other approaches with an F_{macro} score of 0.656.

Overall, we observed that the ensemble models performed well when compared to individual LLMs. Ensembling the models provides additional cues from each individual model, which helps enhance the performance. Furthermore, several variants of the proposed framework outperforms each of the baselines across the evaluated metrics.

Table 8. Samples form the English test dataset where the prediction from the ensemble model (Linear SVC) is accurate, that from the individual LLM is not.

Samples from test set	Ground truth
The Association is also a member of the European Federation for Transport and Environment (EFTE). The Association works closely with other associations that are active within the environmental sector such as the Environment Agency, Europe environment, EFTE and REACH.	B
@snedwan Oh shit We were like one of the most popular bands of the early 90 s and we have some of the best songs	D
But the second half was a completely different story, with the visitors responding with two tries from Andrew Conway and one from Chris Dickson. Conways score came just before the hour mark and gave the visitors a 2017 lead, with Lawrences second coming with just under 10 min to go. However, the Giants hit back with two tries in the dying moments, with Lawrences matchwinning effort on his first start since December making the difference	E

Figure 2 shows the class-wise performance comparison of our best ensemble method (*Linear SVC*) with that of the best baseline (*LR with character-n-grams*) on English and Spanish datasets. For all the classes in both datasets, the macro *F1* score of the proposed method outperforms the baseline macro *F1* scores. Even though the number of parameters for LLMs that we explore are not huge, our proposed ensemble approach performed very well on text generated using the large model with 175B parameters (text-davinci-003).

Tables 8 and 9 show a few samples from the test data for English and Spanish, respectively. In these samples, we demonstrate that while no individual LLM predict the ground truth label correctly, the ensemble Linear SVC classifier predicts the correct label. We also show the ground truth label associated with each sample.

Table 9. Samples form the Spanish test dataset where the prediction from the ensemble model (Linear SVC) is accurate, that from the individual LLM is not.

Samples from test set	Ground truth
La atención recibida por los responsables del Hotel siempre ha sido excepcional, así como la limpieza, calidad de los alimentos En definitiva un 10!	B
Hey seguidores! Sigan a @SoyElHazMeReir en Instagram para ver más contenido: ht t.co/QqJW3YhQ	C
Gracias a la actualización de Internet Explorer, el contenido de estas páginas se encuentra protegido por la ley de derechos de autor,	E

6 Conclusion

In this paper, we explored generative language model attribution for English and Spanish languages. We proposed an ensemble neural architecture where the probabilities of individual LLMs are concatenated and passed as input to machine learning models. Each of the variants of the proposed ensemble approach outperformed several traditional machine learning baselines and the individual LLMs for both languages. Our model results in macro F_{macro} scores of 63% and 65.6% on English and Spanish data, respectively, outperforming other baseline approaches. Our analysis showed that our proposed approach is also effective at classifying the samples that are generated using LLMs with large number of parameters. Our approach also performs well for out-of-domain themes since themes in the test dataset were different from the training dataset.Directions for future work include developing a multi-task approach for generative language model attribution as well as exploring other multilingual datasets.

References

1. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
2. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
3. Crothers, E., Japkowicz, N., Viktor, H.: Machine generated text: a comprehensive survey of threat models and detection methods. arXiv preprint arXiv:2210.07321 (2022)
4. Deng, Z., Gao, H., Miao, Y., Zhang, H.: Efficient detection of LLM-generated texts with a Bayesian surrogate model. arXiv preprint arXiv:2305.16617 (2023)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
6. Gehrmann, S., Strobelt, H., Rush, A.M.: GLTR: statistical detection and visualization of generated text. arXiv preprint arXiv:1906.04043 (2019)
7. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=XPZIaotutsD
8. Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650 (2019)
9. Jawahar, G., Abdul-Mageed, M., Lakshmanan, L.V.: Automatic detection of machine generated text: a critical survey. arXiv preprint arXiv:2011.01314 (2020)
10. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
11. Li, B., Weng, Y., Song, Q., Deng, H.: Artificial text detection with multiple training strategies. arXiv preprint arXiv:2212.05194 (2022)
12. Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H., Gichoya, J.W.: Ethics of large language models in medicine and medical research. Lancet Digit. Health **5**, e333–e335 (2023)
13. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692
14. Massarelli, L., et al.: How decoding strategies affect the verifiability of generated text. arXiv preprint arXiv:1911.03587 (2019)
15. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305 (2023)

16. Mitrović, S., Andreoletti, D., Ayoub, O.: ChatGPT or human? Detect and explain. explaining decisions of machine learning model for detecting short ChatGPT-generated text. arXiv preprint arXiv:2301.13852 (2023)
17. Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156 (2023)
18. Sarvazyan, A.M., González, J.Á., Franco Salvador, M., Rangel, F., Chulvi, B., Rosso, P.: AuTexTification: automatic text identification. In: Procesamiento del Lenguaje Natural. Jaén, Spain (2023)
19. Scao, T.L., et al.: Bloom: a 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
20. Sun, Z.: A short survey of viewing large language models in legal aspect. arXiv preprint arXiv:2303.09136 (2023)
21. Uchendu, A., Le, T., Shu, K., Lee, D.: Authorship attribution for neural text generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8384–8395 (2020)
22. Wolf, T., 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, pp. 38–45 (2020)
23. Wu, S., et al.: BloombergGPT: a large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
24. Zhou, J.T., Tsang, I.W., Pan, S.J., Tan, M.: Heterogeneous domain adaptation for multiple classes. In: Artificial Intelligence and Statistics, pp. 1095–1103. PMLR (2014)