SINAI at IberAuTexTification in IberLEF 2024: Perplexity Metrics and Text Features for Classifying Automatically Generated Text.

César Espin-Riofrio[1,*], Jenny Ortiz-Zambrano[1] and Arturo Montejo-Ráez[2]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador
[2]University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

The task of identifying automatically generated text is very important because of the great advances in automatic generation models. It is a bigger challenge to identify it when dealing with texts in different languages. In this paper, we describe our proposed method for IberAuTexTification subtask 1 in IberLEF 2024: Human or Generated. We use several supervised learning classification methods and an ensemble of them, to train with features extracted from texts such as perplexity, phrasing, and frequency of textual connectors and punctuation marks. Our results, evaluating metrics in the training phase, were promising, but were not good in predicting the final test. We will continue experimenting to improve the model.

Keywords

Human or Generated, Automatically generated text, Text classification, Natural Language Processing

1. Introduction

In recent years, there have been significant advances in automatic text generation models, which mimic human language and can produce compelling text that may confuse readers. The IberAuTexTification task [1] of the IberLEF 2024 evaluation forum invites participants to investigate methods and algorithms for the automatic identification of artificially generated content in various languages.

Automatic text generators have become valuable tools because of their ability to produce text that closely resembles human-generated text. Despite their usefulness, overuse of these tools raises concerns about how they might affect the development of fundamental skills such as creativity and self-judgment [2].

Detecting whether a text was written by a human or automatically generated is an active challenge in the field of Natural Language Processing (NLP) and artificial intelligence. Leading technology companies such as Google, OpenAI and others are working on the development of artificial intelligence-generated text detectors.

The unrestricted deployment of Large Language Models (LLMs) could result in harmful outcomes like plagiarism, the creation of fake news, spamming, and similar issues. Reliable detection of AI-generated text can be critical to ensure responsible use of LLMs [3].

Human texts tend to reflect a unique and consistent style in terms of word choice, sentence structure and lexical preferences. Stylometric models can analyze these characteristics to identify differences between human and automatically generated texts.

They may have statistical anomalies, such as unnatural word patterns or unusual word distributions. Statistical techniques can be used to compare these features with those of human texts [4]. Machine learning classification models can be trained on a labeled dataset with common algorithms such as Support Vector Machines (SVM), Random Forests, among others.

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
cesar.espinr@ug.edu.ec (C. Espin-Riofrio); jenny.ortizz@ug.edu.ec (J. Ortiz-Zambrano); amontejo@ujaen.es (A. Montejo-Ráez)
0000-0001-8864-756X (C. Espin-Riofrio); 0000-0001-6708-4470 (J. Ortiz-Zambrano); 0000-0002-8643-2714 (A. Montejo-Ráez)

Automatically generated texts may lack thematic coherence or cohesion compared to human texts. Algorithms can be used to measure textual cohesion and detect discrepancies. An analysis of the use of punctuation marks and word connectors could then be performed.

Perplexity is a fundamental tool for measuring the quality and predictive ability of language models in machine learning, where lower values indicate better model performance [5]. It can provide useful information as part of a suite of detection techniques, perhaps not as a definitive or stand-alone method for distinguishing between automatically generated text and human text. It may be most effective when combined with other approaches, such as style analysis and statistical analysis, to improve accuracy in detecting automatically generated text.

2. Related work

The following are some recent proposals for the classification of texts written by humans or by a text generator, which use different text characteristics

A study to assess the ability of non-experts to distinguish between human- and machine-written texts (GPT-2 [6] and GPT-3 [7]) in three domains (stories, news, and recipes) was conducted in the research of [8]. On the other hand, in [9] they test the ability of Large Language Models to fool stylometric authorship attribution approaches, using GPT-3 to generate texts that mimic the authors' style.

In [10], they examine the originality of content created by ChatGPT, showing that this model has significant potential to produce exceptional text that eludes accurate detection by plagiarism detection software, while in [4] they explore the differences in how effectively humans and automated systems can recognize text generated by TGM.

Researchers such as [11] have used contextual word embeddings of a BERT model to assess the quality of generated text. In [12] a fine tuning of a predefined model, using as features the initial token embeddings of all layers of the BERT-based transformer model, is performed to identify human-written or automatically generated text.

Giant Language Model Test Room (GLTR), a computer-generated text detector based on textual statistics, was developed by researchers in [13]. The main idea is that generated texts tend to adhere to restricted linguistic patterns.

In [14] they propose an algorithm that uses stylometric cues to detect AI-generated tweets, with models to quantify stylometric changes between human tweets and AI-generated tweets in two key areas: distinguishing between human tweets and AI-generated tweets, and determining when an AI starts generating tweets on Twitter and at what time it does so. Their experiments indicate that stylometric features improve the effectiveness of advanced AI-generated text detectors. The limitations of stylometry for detecting machine-generated fake news are studied by [15].

The Random Forest, MultiLayer Perceptron, XG Boost classification models, are frequently used in binary text classification tasks [16][17].

Perplexity is used as a method for discrimination of similar languages [18], towards few-shot fact-checking via perplexity [19], linguistic investigation on neural language models perplexity [20]. In [21] they analyze the characteristics of the translations obtained by five state-of-the-art Spanish-Basque translation systems, measuring lexical diversity and density, and perplexity.

3. Method

In this section we present the method used in the experimentation.

We used classification methods, Random Forest, MultiLayer Perceptron, XGB Boost and an ensemble of them, to train with the following features extracted from the texts:

- **phraseological** (sentences per paragraph and their mean, standard deviation, and variance; average sentence and word length per paragraph; average sentence length difference per paragraph; words per paragraph; lexical diversity).

- **word connectors** (frequency of use per text according to the list of most used connectors for each language).
- **punctuation marks** (frequency of use of punctuation marks per text).
- **perplexity metric** (measures the efficiency of a language model in generating text according to the probability of the next word, specific to each language).

3.1. Data

The dataset [22] used was provided by the organizers of the IberAuTexTification in IberLEF 2024 event for subtask 1: Human or Generated.

The dataset contains texts from a variety of domains such as essays, news, social media (tweets, forums, dialogues), Wikipedia, Wikihow, among others. These texts were generated by LLMs such as GPT-3.5, GPT-4, LLaMA, Coral, Command, Falcon, MPT, among others, using the TextMachina tool. They also cover the main languages of the Iberian Peninsula: Spanish, Catalan, Basque, Galician, Portuguese and English (in Gibraltar).

3.2. Pre-processing

First, we encode the labels "generated" and "human" of the target column "label" with 0 and 1 respectively

Next, we perform the automatic language detection of all the texts. For this purpose, the fasttext tool [23] was used, which is very accurate in identifying the language of a text. Very few texts had to be identified manually. The languages were labeled in a new column of the dataset.

Now, there was the need to encode them to use them as part of our features. This was done by means of one hot encoding that separates each language independently indicating with zero or one which one it is. This technique was used because the new column that identifies the languages corresponds to categorical variables, which, if only coded, could cause bias problems during training.

3.3. Text features

As already mentioned, the characteristics extracted were phraseological, perplexity, frequency of textual connectors and punctuation marks, which we describe as follows.

3.3.1. Perplexity

To calculate text perplexity we refer to that proposed in [24]. A pre-trained text generation model is needed to analyze the probability of the next word for each text. We use the language-specific GPT-2 models.

Table 1
GPT-2 models for each language

Language	GPT-2 model
Spanish	DeepESP/gpt2-spanish-medium
English	openai-community/gpt2-large
Catalan	ClassCat/gpt2-small-catalan-v2
Portuguese	egonrp/gpt2-small-portuguese
Basque	HiTZ/gpt2-eus-euscrawl
Galician	fpuentes/gpt2-galician

All of the above models can be obtained from the Hugging Face model repository, [25].

3.3.2. Phraseological

The analysis of phraseological features helps us to identify style-related differences between human and automatically generated texts.

Several phraseological characteristics were extracted from the texts:

- average number of sentences per paragraph
- standard deviation of the number of sentences per paragraph
- variance of number of sentences per paragraph
- average sentence length variance per paragraph
- word length per paragraph
- lexical diversity

The lexical diversity of the texts is calculated by measuring the variety of unique words in relation to the total number of words in the text. The other phraseological characteristics are calculated in relation to the frequency of words per sentence and the number of sentences per paragraph to obtain means, averages, standard deviation and variance.

3.3.3. Frequency of textual connectors

We analyzed the frequency of use of textual connectors in the text. For this, we manually obtained the most commonly used textual connectors in each language identified in the dataset, extracting them from the main sites that list and analyze their use in the writing of each language, examples are shown in Table 2. Then, we detect their frequency of use in each text of the dataset according to their corresponding language

Table 2
Examples of textual connectors for each language

Language	Textual connectors
Spanish	a continuación, de hecho, como, además, ...
English	actually, also, because, however, ...
Catalan	a fi de, al contrari, com a exemple, d'entrada, ...
Portuguese	a fim de, a ñao ser que, contudo, é o caso de, ...
Basque	aitzitik, alabaina, baina, baita ere, ...
Galician	á diferenza de, ademais, aínda así, certamente, ...

3.3.4. Frequency of punctuation marks

The frequency of use of punctuation marks in the text may give clues as to when a text has been generated automatically. For this purpose, we rely on lists of punctuation marks commonly used in all languages.

Examples of punctuation marks are: ¡ ! ¿ ? ( ) [ ] " : ; , .

Punctuation marks and phraseological features are calculated independently of the language of the text. In contrast, perplexity and textual connectors are calculated according to each language of the texts in the dataset.

4. Experiment

The implementation is based on the Random Forest, MultiLayer Perceptron, XG Boost classification models, and an ensemble of them. The Voting Classification technique was used to create the ensembled model.

We used the Cross Validation technique to evaluate the models during the training phase.

Finally, ablation tests were performed with the set of extracted features to check their influence on the model performance and to determine with which combination of them the best prediction results are obtained.

5. Results

As mentioned, the results during training were very good. Table 3 shows the ablation tests for the ensemble model, using or not using each feature, and all together.

Table 3
Ablation tests with ensemble model

	Only			**Without**		
	Precision	Recall	F1	Precision	Recall	F1
phraseological	0.814027	0.814025	0.813654	0.858835	0.858524	0.858208
punctuation marks	0.742604	0.741759	0.739986	0.877422	0.877445	0.877432
word connectors	0.751320	0.751516	0.751384	0.883920	0.883965	0.883912
perplexity	0.722775	0.721105	0.721438	0.862849	0.862718	0.862763
All	0.889590	0.889573	**0.889581**			

"Only": using only that feature.
"Without": using all features except that.

Fig.1 shows the metrics of each model when all the features were used together. the results obtained for accuracy, precision, recall and f1-score are shown.

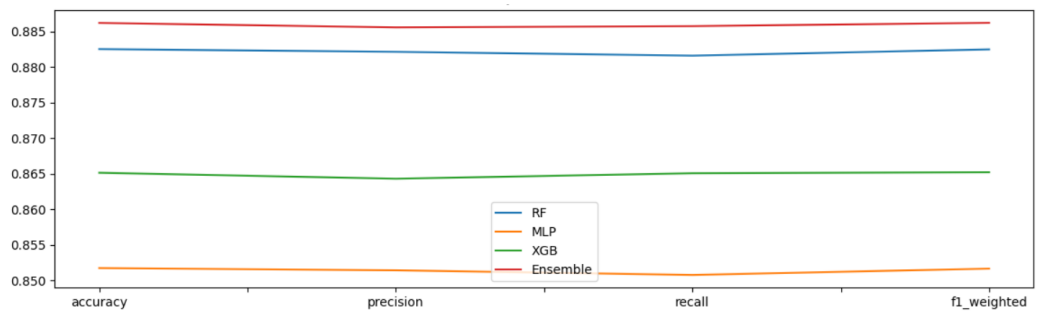

Figure 1: Model evaluation during training.

Table 4 shows our results (**team sinai**) obtained in the ranking by teams participating in the task, which were not as good as expected.

All the details of the IberAuTexTification workshop, including information on the different participants and the results obtained, are available in the official overview [26].

6. Conclusions

In this study, a text classification task was carried out in order to distinguish between texts generated by text generation models (TGM) and those produced by human beings. we use perplexity features, phraseological features, textual connectors and punctuation marks to train the Random Forest, MultiLayer Perceptron, XG Boost algorithms and a ensembled model from them.

The results indicate that the model performs moderately well, achieving an F1 score of 0.6040 for the test texts evaluated. They were not as expected, since during the training evaluation we obtained F1

Table 4
IberAuTexTification subtask1 competition ranking

Rank	Team	Macro-F1
1	jor_isa_uc3m	0.8050
2	gmc_fosunlp	0.7663
3	telescope_team	0.7579
4	iimasNLP	0.7188
5	gmc_fosunlp	0.7155
...	...	...
35	**sinai**	**0.6040**
...	...	...
53	paporomerol	0.5629
54	Yano	0.5608

scores of 0.889581 which generated a lot of expectation. This is probably because the texts proposed in the test dataset are from different domains than the dataset with which the training was performed.

We can distinguish from the ablation tests that phraseological features alone have a greater influence on model performance than when not used, while perplexity alone has little influence, proving that together with other features it does produce better results.

On the other hand, Random Forest followed by XG Boost are the classifiers with which, in all tests, the best results were achieved.

Overall, the result obtained indicates that with the features evaluated and the classifier models, it is possible to differentiate between machine-generated text and human-written text with some reliability, although it is evident in our case that there is much room for improvement. In the future, we will explore techniques that help classify text from different domains as well as using additional features or other learning models.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government.

References

[1] A. M. Sarvazyan, J. Á. González, F. Rangel, P. Rosso, M. Franco-Salvador, Overview of iberautextification at iberlef 2024: Detection and attribution of machine-generated text on languages of the iberian peninsula, Procesamiento del Lenguaje Natural 73 (2024).

[2] C. Espin-Riofrio, R. E. Fajardo, F. J. O. Serrano, T. P. Guaraca, R. C. Encalada, Dataset de textos en español de ecuador con cuatro versiones reescritas por gpt para tareas de identificación de texto generado automáticamente, Polo del Conocimiento 9 (2024) 998–1010. URL: https://www.polodelconocimiento.com/ojs/index.php/es/article/view/6570https://www.polodelconocimiento.com/ojs/index.php/es/article/view/6570. doi:10.23857/pc.v9i2.6570.

[3] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can ai-generated text be reliably detected?, arXiv preprint arXiv:2303.11156 (2023).

[4] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1808–1822. URL: https://aclanthology.org/2020.acl-main.164. doi:10.18653/v1/2020.acl-main.164.

[5] F. Jelinek, R. L. Mercer, L. R. Bahl, J. K. Baker, Perplexity—a measure of the difficulty of speech recognition tasks, The Journal of the Acoustical Society of America 62 (1977) S63–S63.
[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[7] 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.
[8] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, All that's' human'is not gold: Evaluating human evaluation of generated text, arXiv preprint arXiv:2107.00061 (2021).
[9] S. Rebora, et al., Gpt-3 vs. delta. applying stylometry to large language models, in: La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, 2023, pp. 292–297.
[10] M. Khalil, E. Er, Will chatgpt get you caught? rethinking of plagiarism detection, 2023. arXiv:2302.04335.
[11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. arXiv:1904.09675.
[12] C. Espin-Riofrio, J. Ortiz-Zambrano, A. Montejo-Ráez, Sinai at autextification in iberlef 2023: Combining all layer embeddings for automatically generated texts. (2023).
[13] S. Gehrmann, H. Strobelt, A. Rush, GLTR: Statistical detection and visualization of generated text, in: M. R. Costa-jussà, E. Alfonseca (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 111–116. URL: https://aclanthology.org/P19-3019. doi:10.18653/v1/P19-3019.
[14] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, H. Liu, Stylometric detection of ai-generated text in twitter timelines, arXiv preprint arXiv:2303.03697 (2023).
[15] T. Schuster, R. Schuster, D. J. Shah, R. Barzilay, The limitations of stylometry for detecting machine-generated fake news, Computational Linguistics 46 (2020) 499–510.
[16] K. Salehin, M. K. Alam, M. A. Nabi, F. Ahmed, F. B. Ashraf, A comparative study of different text classification approaches for bangla news classification, in: 2021 24th International Conference on Computer and Information Technology (ICCIT), IEEE, 2021, pp. 1–6.
[17] A. Occhipinti, L. Rogers, C. Angione, A pipeline and comparative study of 12 machine learning models for text classification, Expert systems with applications 201 (2022) 117193.
[18] P. Gamallo, J. R. P. Campos, I. Alegria, A perplexity-based method for similar languages discrimination, in: Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial), 2017, pp. 109–114.
[19] N. Lee, Y. Bang, A. Madotto, M. Khabsa, P. Fung, Towards few-shot fact-checking via perplexity, arXiv preprint arXiv:2103.09535 (2021).
[20] A. Miaschi, D. Brunato, F. Dell'Orletta, G. Venturi, What makes my model perplexed? a linguistic investigation on neural language models perplexity, in: Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, 2021, pp. 40–47.
[21] N. Aranberri, Can translationese features help users select an mt system for post-editing? (2020).
[22] J. González, A. M. Sarvazyan, M. Franco-Salvador, F. M. Rangel, P. Rosso, Iberautextification (test dataset), 2024. URL: https://doi.org/10.5281/zenodo.11034382. doi:10.5281/zenodo.11034382.
[23] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).
[24] HuggingFace, Perplexity of fixed-length models, https://huggingface.co/docs/transformers/perplexity#perplexity-of-fixed-length-models, 2024. Accessed: 2024-07-02.
[25] HuggingFace, Trending gpt-2 models, https://huggingface.co/models?sort=trending&search=gpt2, 2024. Accessed: 2024-07-02.
[26] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages

Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.