The influence of the perplexity score in the detection of machine-generated texts

Alberto J. Gutiérrez-Megías and **L. Alfonso Ureña-López** and **Eugenio Martínez-Cámara**
SINAI Research Group, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén (Spain)
{agmegias, laurena, emcamara}@ujaen.es

Abstract

The high performance of large language models (LLM) generating natural language represents a real threat, since they can be leveraged to generate any kind of deceptive content. Since there are still disparities among the language generated by machines and the human language, we claim that perplexity may be used as classification signal to discern between machine and human text. We propose a classification model based on XLM-RoBERTa, and we evaluate it on the M4 dataset. The results show that the perplexity score is useful for the identification of machine generated text, but it is constrained by the differences among the LLMs used in the training and test sets.

1 Introduction

Large language models (LLMs) present a large number of capabilities, ranging from text summarization and information extraction to text paraphrasing (Wei et al., 2022). One of those abilities is text generation, which is approaching to the human written performance (Li et al., 2021; Minaee et al., 2024). However, they also present some pitfalls that can lead to privacy and security leaks. For instance, the tendency to hallucinate of LLMs may lead to privacy violations by exposing sensitive data (Ji et al., 2023). Likewise, the generative capacity of LLMs is an extremely positive skill for many applications, but it may be used to generate deceptive and malicious content, which can be used as a source of security leaks (Jawahar et al., 2020; Peng et al., 2018; Das et al., 2024). Hence, we need the automatic identification of machine generated text to warm about it to the readers.

We can consider the language generated by each person that follows a particular probability distribution. Although, the small nuances among the use of language of each person, the spoken and written language by humans follow a common probability distribution. Similarly, the language generated by LLMs follows a specific probability distribution, with some disparities between LLMs, but with a large difference with respect to the human language. Perplexity is a measure of uncertainty in the value of a sample from a discrete probability distribution (Rosenfeld et al., 1996). Accordingly, a low value of perplexity means a reduced uncertainty score that the sample is drawn from a probability distribution, otherwise it is likely that the sample does not belong to the distribution. Hence, perplexity can be used to discern whether a span of text follows the probability distribution of the language usually generated by a LLM or by a human.

In this work, we claim that perplexity can be used as a classification signal for identifying span of text generated by machines, with the aim of warming readers and protecting them from deceptive content. We thus propose a classification system built upon the XLM-RoBERTa language model (Conneau et al., 2019), whose input are the word embeddings vectors of each input token and the perplexity score of the input text.

We evaluate the classification model on the M4 dataset (Wang et al., 2024b) used in the task 8 of SemEval (Wang et al., 2024a). Moreover, we analyze whether there is any influence in the nature of the LLM used to calculate the perplexity score and the one used to generate the evaluation texts.

The results show that the perplexity is a useful signal to identify machine-generated texts, but it is limited to a small difference among the probability distribution of the LLM used to calculate its score and the one used to generate the text to classify.

This work is organized as follows: next section highlights the most salient related works. Section 3 justifies the use of perplexity as classification signal. Our proposal is described in Section 4, and the experimental framework in Section 5. Then, we analyze the results in Section 6, and we remark the main conclusions in Section 7.

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 80–85
July 29–30, 2024.

2 Related work

LLMs are able to generate text very similar to what a human can do. Accordingly, differentiating a machine-written text from a human one is very challenging (Crothers et al., 2023). The automatic detection of these kinds of text is crucial to security scenarios like phishing, fake news, identity fraud, and others. Powerful models are open to use by anyone with the capability to connect to the internet, such as those ones in Hugging Face[1]. This facility for the user to be able to generate any type of text with hardly any resources demonstrates the importance of obtaining a system that can differentiate when a text is artificially generated.

The need of recognizing machine or artificial intelligence (AI) generated text comes from the first uses of GROVER (Zellers et al., 2019) for the generation of propaganda. Since that moment several models and methodologies have been published to detect this automatic generated text, because humans struggle at it (Dugan et al., 2023).

We mainly find two approaches to face up the challenge of detecting AI generated text. On the one hand, the proposals based on used of linguistic features, as for instance TF-IDF (Fröhling and Zubiaga, 2021) or the use of fluency features as the Flesch score (Crothers et al., 2022). On the other hand, the works ground in the use of language models. For instance, in (Rodriguez et al., 2022), the authors fine-tuned a RoBERTa model to detect GPT-2 generated texts. Likewise, in (Kushnareva et al., 2021) is shown that features derived from BERT outperform linguistic and other features stemmed from other neural models.

The literature of machine generated text detectors is wide (Crothers et al., 2023; Valiaiev, 2024), but as far as we know, perplexity has not been used yet as feature to guide the identification of machine generated text. In this paper, we claim to use perplexity as a classification signal, and it shows to give a strong performance as we show in the subsequent sections.

3 Perplexity as feature

Perplexity is a metric from information theory that indicates how well a probability distribution or model predicts a given sample. Its usefulness resides in facilitating the comparison of various probability models (Jelinek et al., 1977). A low value of perplexity means that a sample may be derived from the probability distribution, since there is a low value of uncertainly, otherwise the perplexity value is large.

Perplexity is usually lower in texts generated by AI and their texts rather express feelings and use unusual words. Crothers et al. (2023) show a difference in performance between perplexity-based and machine learning-based classification, the latter being better than perplexity-based classification. Consequently, the use of both parameters, text, and perplexity, in training a classifier may be interesting to study in this task, demonstrating that the use of perplexity in texts generated by the LLM itself results in highly accurate results.

The perplexity of human-generated text tends to be higher than that of machine-generated text according to (Mitrović et al., 2023), because the perplexity is calculated according to a specific LLM, which generates language that follows a different probability distribution than the human language. Hence, we calculate the perplexity score of the dataset text that we will use for training and evaluation of the system (see section 5.1). To calculate the perplexity score we used the Language Model Perplexity (LM-PPL) python library.[2] The LM-PPL computes an ordinary perplexity for recurrent LMs such as GPT3 (Brown et al., 2020). We calculate the perplexity score of each instance using the GPT2 language model (Radford et al., 2019). Table 1 shows the perplexity score of human language and the text generated by several LLMs. As the table shows, there is a large disparity among the perplexity score of human language and the language automatically generated. Therefore, we can use perplexity as an additional feature to classify machine generated text.

The perplexity *PP* of a discrete probability distribution *p* is a widely used concept in information theory, where *H(p)* is the entropy of the distribution, and *x* ranges over the events.

$$PP(p) = 2^{H(p)} = 2^{p(x)\log_2 p(x)} = log \quad (1)$$

4 Machine-detection system

We have developed a fine-tuning classification model based on XLM-RoBERTa for differentiating text authorship. This Machine-detection (MD) system uses the text and the perplexity associated with

[1] https://huggingface.co/models

[2] https://pypi.org/project/lmppl/

Generator Model	Mean Perplexity
Human	34.1865
ChatGPT	12.1334
Cohere	11.3244
Davinci	22.6191
Bloomz	30.1235
Dolly	18.9728

Table 1: Comparison of perplexity means of different models including the human-written text.

each text as input parameters. The perplexity value has been calculated using LM-PPL with the GPT2 model as a reference.

To fuse the two feature sets, we use the Multimodal Toolkit library, which offers several fusion methods. In this case, we have selected a specific approach that involves multi-layer perceptron (MLP) partitioning for categorical and numerical features. Subsequently, the output of the transformer is concatenated with processed numerical and categorical features before reaching the final classifier. Once it reaches the classification head, the system is trained. To optimize this training, we performed a hyperparameter optimization (see Section 5.2). We depict the system in figure 1.

5 Experimental framework

We have developed a training system including data from all LLMs as a baseline for our experimental framework. One training has been conducted using perplexity and the other without it. We have also assessed the performance of our proposal when the training and test texts have been generated using the same LLM, and we compare them when that difference is not done. This proves that the use of perplexity improves the performance of the system when the model is trained and evaluated using the machine-generated text by the same LLM.

These two baselines allow us to compare them with our proposed system, demonstrating that the use of perplexity improves the hit rate in identifying the authorship of the text when training and predicting the generated text with the same linguistic model.

Baseline one - fine-tuning The system without perplexity value is a fine-tuning using the XLM-RoBERTa-Large, trained with a balanced dataset where the machine-generated text used is comprised of all the texts of the LLMs.

Baseline two - fine-tuning and perplexity This baseline is similar to the previous system but with the addition of perplexity. The same dataset is used in this system. This system is the same as we propose, the only difference is the training data used.

5.1 Dataset

The M4 dataset (Wang et al., 2024b) consists of 71,027 instances assigned to training, 3,000 instances for development, and 18,000 instances designated for final predictive testing. All data in this dataset are in English. Each instance is characterized by its textual content and the specific model for its generation. Non-machine-generated instances are indicated by the label *human*. Possible generating models include *ChatGPT*, *Cohere*, *Davinci*, *Bloomz* and *Dolly*, each representing 16.6% of the dataset. This distribution results in an unbalanced binary task classification since more than 80% of the instances consist of machine-generated text.

An additional dataset providing human-generated text from the SemEval 2024 competition was integrated to ensure a balanced representation within the dataset, tailored to this specific classification task. For the dataset used to train our proposal, the machine detection (MD) System, the dataset was split, each comprising exclusively instances generated by one of the five LLMs contained in the dataset and human texts.

The training datasets to create the models capable of differentiating between text and machine of a specific LLM is composed only of text generated by that LLM and human text so that the dataset is balanced. This process has been done five times, once for each LLM in the dataset.

5.2 Model detection training

For systems involving fine-tuning, we used Optuna (Akiba et al., 2019), a hyperparameter optimization software framework. The fine-tuning process consisted of investigating these values, with all systems using the same optimization parameters. To perform these searches, we used a development set consisting of the 3000 instances described above. During the final model training phase, we merged this development set with the training set to increase the quality of the training.

To ensure the reproducibility of the experiments we present the values explored for optimization. The hyperparameter values for Epochs are [8,16], Learning Rate [5e-6, 5e-5], Weight Decay [1e-12, 1e-1] and Adam Epsilon [1e-10, 1e-6].

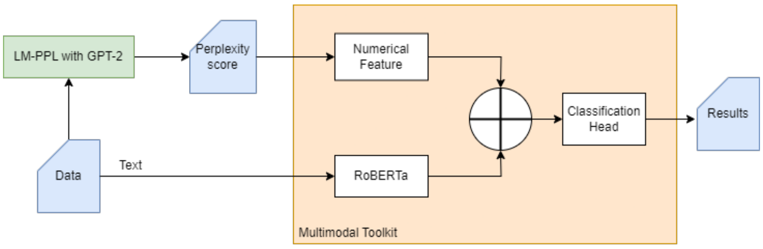

Figure 1: Structure of the Machine-detection System using perplexity and text for its development.

System	Precision	Recall	F1
Baseline One	0.9507	0.7309	0.7903
Baseline Two	0.8670	0.8624	0.8619
MD System - ChatGPT	0.9272	0.9148	0.9142
MD System - Cohere	0.8725	0.8717	0.8714
MD System - Davinci	0.9361	0.7581	0.7432
MD System - Bloomz	0.9996	0.9996	0.9996
MD System - Dolly	0.8015	0.671	0.6310

Table 2: Final results of the experiments.

With Epochs 8, Learning Rate 1.64E-05, Weight Decay 9.41E-08 and Adam Epsilon 5.51E-07 being the final values of the optimised hyperparameters for all experiments.

6 Results and discussion

As shown in table 2 The macro-F1 score shows a decrease compared to that of MD System in most cases. In particular, the recognition of textual authority improves significantly when the system is trained and predicted with machine-generated text from the same LLM system.

The disparity between Baseline One and Two lies in the macro-F1 score demonstrating the improvement of the system when perplexity is added to the training. While Baseline One exhibits superior precision in generating machine text, Baseline Two demonstrates a broader efficacy. Notably, Baseline Two excels in discerning between human and machine-generated text owing to its balanced consideration of the macro-F1 score for both categories.

The results of *Bloomz* have obtained a macro-F1 score of more than 0.90, almost perfect. In contrast, the *Dolly* shows lower results than Baseline One and Two. The analysis reveals no significant correlation between the average perplexity of a model. The *Bloomz* has an average perplexity similar to that of a text written by a human being, but its results are much higher.

Using the methodologies defined in this study, evidence emerges for the effectiveness of using perplexity in conjunction with textual features to classify authority. On the other hand, in cases where there is certainty about the uniformity of the LLM model across machine-generated text, the effectiveness of such classification depends on the models used and the methodologies employed to calculate the perplexity score.

Our hypothesis holds in most cases. With the systems that have been trained with the *ChatGPT*, *Cohere*, and *Bloomz* models we obtain a macro-f1 superior to Baseline Two, being remarkable improvement where the same LLM models are used to train and evaluate the experiments. Even, in systems such as the one used by *Davinci* where the macro-f1 is lower than Baseline Two, we can see an improvement in accuracy.

7 Conclusions

The results obtained have shown that the performance obtained has been improved for most of the LLM models that have been worked with. This shows that as long as the same LLM generates the machine-generated data the proposed system using perplexity and text can with a high probability of success differentiate between whether a text is machine-generated or human-generated.

It is also worth noting the difference in the results between the baselines exposed. This also proves that the additional information on the perplexity of each text is useful information for the authority recognition of the generated text, even if it has been trained by different LLMs.

Following the positive results obtained in MD System, our next objective will be to classify texts independently of their origin. For this purpose, we will apply the same methodology with considerable modifications. Such modifications may include the integration of a new model to calculate text perplexity or the use of several models to generate a vector of perplexities.

Acknowledgements

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146- C21) and FedDAP (PID2020-116118GA-I00) funded by MCIN/AEI/ 10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR".

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Evan Crothers, Nathalie Japkowicz, Herna Viktor, and Paula Branco. 2022. Adversarial robustness of neural-statistical features in detection of generative transformers. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *arXiv preprint arXiv:2402.00888*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771.

Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7:e443.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2020. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language model for text generation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4492–4499. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Tianrui Peng, Ian Harris, and Yuki Sawa. 2018. Detecting phishing attacks using natural language processing and machine learning. In *2018 ieee 12th international conference on semantic computing (icsc)*, pages 300–301. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, Seattle, United States. Association for Computational Linguistics.

Ronald Rosenfeld et al. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer speech and language*, 10(3):187.

Dmytro Valiaiev. 2024. Detection of machine-generated text: Literature survey. *arXiv preprint arXiv:2402.01642*.

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. 2024a. Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. *Defending against neural fake news*. Curran Associates Inc., Red Hook, NY, USA.