UDC 004.93 **doi: 10.26906/SUNZ.2026.2.153**

Olena Peredrii, Oleksii Gorokhovatskyi

Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine

THE EXPLAINABILITY OF SHALLOW AI-GENERATED TEXT CLASSIFICATION MODELS VIA PARTS REMOVING

Abstract. In this paper, we address the explainability problem for the ANNs' classification of AI-generated and human-written text chunks in Ukrainian texts in the IT domain. The objective is to investigate whether the perturbation-based modifications of text chunks that include the removal of sentences, words, and word combinations may be helpful in searching for explanations. We used five shallow ANN models (with an average accuracy of about 0.88) and tested them on a sample of the document containing human-written text and AI-generated fragments generated with GPT-5, Gemini 2.5 Flash, and Claude Sonnet 4.5. The experimental modeling showed that it is not easy to find a single sentence or word that can flip the classification result. We have proposed an explainability index that measures the total influence of all perturbed samples on the classification result, accounting for the fact that short perturbations are more valuable.

Keywords: explainability, black-box, shallow ANN, perturbation, AI-generated content, human-written content, text chunk, text classification, explainability index.

Introduction

The integration of artificial intelligence (AI) systems into essential human domains, such as healthcare, finance, education, and identification systems, as well as autonomous decision-making, requires a high level of explainability for their processes and outcomes. The inability to trace how a specific input leads to an output poses significant model usage risks, which are relevant both for modern transformer-based architectures like LLMs and for traditional AI methods and tools like artificial neural networks [1, 2].

The paper analyzes the application of artificial neural networks (ANNs) for classifying AI-generated and human-written text chunks (e.g., paragraphs) in Ukrainian. We do know that the application of AI detectors is commonly recommended to be limited in educational environments, so our primary goal is mostly research-oriented. In particular, we are interested in identifying details that may help explain why the model predicts the specific class and which parts of the text are primarily responsible.

1. Literature review

Traditional AI models used for classification or regression rely on explicit algorithms and simpler internal structures compared to modern large language models. These systems typically use fixed inputs to predict a specific label, and their explainability methods focus on revealing the reasoning behind these discrete choices.

There are different classifications of existing Explainable AI (XAI) methods; the most popular is whether the weights of the initial black-box model are required. Model-agnostic methods treat the original model as a black-box entity without access to its weights; on the other hand, model-specific methods leverage access to the model's structure and weights [13]. Perturbation-based methods are representatives of model-agnostic approaches that modify the initial signal (numerical features, images, text, etc.) representing the entity being classified and investigate how these changes affect black-box model outputs. According to [12], there are also score-based methods, explanations by simplification, language explanations, and different methods that can fall into multiple of these categories.

XAI explainable methods can also be categorized into local and global [15], where local methods search for explanations based on the particular input, and global methods analyze the behavior of the black-box model as a whole.

There are many different explanation methods already known for convolutional neural networks (CNNs) and image classification problems [3]. They include the CAM/Grad-CAM family [4, 5], methods based on the weighted sum of feature maps from the final convolutional layer, or on gradients flowing into it. The following researches about Integrated Gradients [6] (calculates the integral of gradients along a straight-line path from a "baseline" input to the actual one) and DeepLift [7] (compares the activation of each neuron to a reference state and decomposes the output prediction based on these differences) addressed the common issues for the gradient-based methods: the saturation of neurons even for extremely important features of the signal, as well as that gradients provide the information only about some neighborhood around the point that may not reflect the global importance of this feature.

One of the most famous perturbation-based explainability methods is LIME (Local Interpretable Model-agnostic Explanations) [8]. The research asserts that explanations must be interpretable (human-readable) and locally faithful (accurately reflecting how the model behaves in the immediate vicinity of the data point being explained). LIME creates a "neighborhood" of the input signal by randomly perturbing it; afterward, these perturbed samples are fed into the complex "black-box" model to observe how the predictions change. LIME builds the local interpretable model based on how close these samples are to the original input signal. This research provided the first universal "boilerplate" for explaining complex models using simple local approximations.

The application of LIME for text classification problems was investigated in [10]. The paper is focused on the TF-IDF transform and the sampling mechanism, which are the two critical components LIME uses to handle text. By analyzing these, the research determines

if LIME's output is a stable and faithful representation of the complex model being explained. The random removal of separate words was used as a sampling technique; it was shown that LIME provides meaningful explanations for decision trees and linear models.

The main benefit of LIME is its broad applicability to any classifier, making it a convenient tool for proprietary or inaccessible models. However, LIME's process requires repeated evaluations of the target model, which can be computationally intensive, especially for large LLMs or long text sequences. This method can also introduce noise.

The SHAP (SHapley Additive exPlanations) proposed in [9] provides a theoretical framework that unifies several existing explanation methods—including LIME, DeepLIFT, and Layer-Wise Relevance Propagation. SHAP assigns an importance score (within the scope of the particular prediction case) to each input feature, which is considered mathematically fair for distributing the prediction score among all input features.

Model-agnostic techniques are commonly applied to artificial neural networks (ANNs) and tabular data. LIME constructs local surrogate models via input perturbations, whereas SHAP employs game-theoretic methods to equitably attribute feature impact. Although both methods are theoretically robust, SHAP incurs high computational costs, and LIME may exhibit instability due to random sampling.

The research [11] describes the investigation of how Integrated Gradients (IG) can be used to provide word-level explainability for text classification models. The BERTAgent was used as a primary deep-learning classifier, and integrated gradients were applied to reveal which specific words contribute to a model's decision. It was concluded that the quality of the output strongly depends on the specific label, the availability of strong (and statistically coherent) semantic content in the data, and whether the classifier captures it. It stated that the application might be useful for extending their theoretical background in classification problems based on uncertain data.

The paper [14] introduces ProtoLens, which searches for explanations in sub-sentence text chunks rather than words, making the results more human-understandable. The model maps text spans to known prototypes and assigns scores and weights to prototypes based on their similarity to the span. The modeling showed that ProtoLens outperforms various existing baseline methods, providing intuitive and detailed explanations. The main limitation of the approach, as mentioned by the authors, is its reliance on the training data, which may be biased.

The HINT (Hierarchical Interpretable Neural Text classifier) is a model that generates the hierarchical representation of label-associated topics of model predictions for word-level, sentence-level, and document-level analysis [15]; the authors claimed that the interpretations by words and phrases are not sufficient. Results of experiments for two datasets confirmed the effectiveness of the HINT, comparable to the SOTA classifiers, and its ability to generate faithful and humans-readable interpretations.

The detection of adversarial attacks on text classifiers based on explainable AI with integrated gradients is proposed in [16]. The core of the method is the gradient-based identification of words that affect the prediction result positively and negatively, with the following replacements of such words with synonyms instead of replacing random words. This method shows an example of how an XAI method could be used both to detect adversarial attacks and to find useful insights about the model's behavior. The main drawbacks of the methods, which are important for us in this paper, are the requirement to use gradients and the necessity to generate high-quality synonyms in the correct form.

As a brief summary of the methods described above, we can conclude the following. There are many successful methods addressing various issues in English, but their applicability, for instance, to Ukrainian, is unclear. Known methods use NLP techniques such as finding appropriate synonyms, which are language-specific. The majority of methods work at the word level, which is suitable for short text chunks, but does not seem to be a very good choice for the paragraphs we are dealing with in the current problem. Finally, most approaches use gradients, and our primary focus is the model-agnostic perturbation-based family.

The contribution of the paper includes:

– the research of the possibility to find the explanations for the black-box shallow ANN models for text chunks classification in Ukrainian using the removal of sentences and n-grams as sampling strategies;

– measuring the explainability index for the particular decision, evaluating the ability to find the complementary text pairs.

2. Ethical considerations

The unreliable performance (mainly in the form of false-positive classification errors) of the majority of modern common-knowledge AI detectors is a known issue. The justified solution we can see frequently involves avoiding AI detectors in education for students or educators, but it is recommended to use improved teaching and process-based assessment methods, oral exams, and discussions, and to educate students on the responsible use of AI tools instead. But it is worth noting that specialized AI detectors are used in academia anyway.

In this and our previous papers, we follow this recommendation: we are building an AI system to detect AI-generated content represented as documents in Ukrainian for scientific purposes. Our main goal is to investigate whether it is possible to implement really honest detection in narrow-field domain, with limited technical resources using shallow ANN architectures [17] and LLM available to everyone (almost) for free, what could be the quality of the solution (for own challenging dataset), and whether it is possible to find the explanations of the particular decision. The application of the tool implemented in this paper is limited to personal use, e.g., testing one's own documents to understand the responsibility, without any official consequences or automatic decisions.

3. Models

We used the ANN models described in detail in [17]. The dataset contained 5167 text chunks (2533 human-written and 2634 generated with the GPT-4o-mini model) in Ukrainian, with the texts spanning bachelor's theses in the IT specialty.

This dataset was used to train various shallow ANN models with grid search, five best models were selected for the further experiments, their architectures and accuracies are presented in Fig. 1. As one can see, these models can predict whether the text chunk was human-written or AI-generated with the rough probability about 0.88, but all models are black-box completely and are not capable of producing any explanations of the decision being made. Class label 0 for all models means "human-written", and label 1 means "AI-generated" content.

It is worth noting that the dataset is challenging. There are many cases where humans cannot correctly determine whether a fragment was generated.

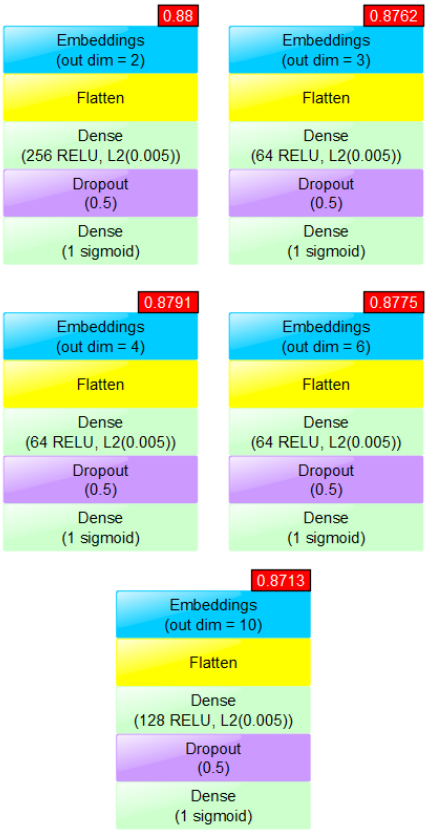

Fig. 1. Architectures of ANN and their accuracies

4. Good explanation

Perturbation-based methods are probably the easiest ones to search for the explanations without using the model's weights and its internal structure. They could be applied to different types of input signals, e.g., images or texts, to evaluate the importance of signal parts for the final decision. There are different types of perturbations: signal part removal, replacement with a predefined constant value, different values, specific modifications to include another signal within the initial one, and so on.

In our previous research [18, 19], we applied the search for complementary parts of the signal to be sure it affects the decision at the required level: we were interested in such a pair of modifications (perturbations) of the initial signal, one of which still preserves the same classification result as the original signal, the second one – changes it. We called such a pair of signals a complementary image pair (CIP) for the image classification problem [18, 19].

The model's initial input is a chunk of text for the problem being investigated. The perturbation of the text segment that leads to the change in classification result is our primary interest.

In this paper, we claim that a good explanation is a piece of input signal that is minimal in size (has minimal length of text for the current problem), changes the classification result, and the presence of this piece is enough to preserve the initial prediction.

So, we can introduce the measure of perturbation size to be (1 – (size of perturbation / size of chunk)) and the influence of this sentence on the result as |etalon prediction – perturbed prediction|. The quantitative measure of the explainability index (EI) of the particular decision by the black-box model may include the normalized sum of explainabilities for all perturbations:

$$EI = \sum_{i=1}^{n} \left(1 - \frac{l_{perturb}}{l_{text}}\right) \left| pred_{et} - pred_{perturb} \right| \qquad (1)$$

where n – is the quantity of perturbations (sentences/n-grams to be removed), $l_{perturb}$ – length of the perturbed fragment, l_{text} – length of the entire chunk, $pred_{et}$ – etalon (initial) prediction for the original chunk, $pred_{perturb}$ – prediction for the perturbed chunk.

The qualitative measure could be a simple flag (yes/no) indicating whether the explanation that flips the classification was found. In this context, it is also clear that it is easier to find an explanation when the initial decision is less confident (for instance, the probability of AI-generated text is 0.75 rather than 0.95). Changing the decision of the etalon model seems to be a good criterion, but just measuring the influence of the parts is easier to implement.

5. Sampling

We have considered sentence, word, and n-gram removal. We also briefly tried sampling based on sentence replacement, but found that mapping the part to be replaced to a similar, grammatically correct human-written chunk is complex.

The first sampling strategy we have tried is the complete removal of the sentence and classification of the text chunk without it. Removing sentences to measure their influence seems like a pretty obvious idea, but it could be tricky, given that the models we use were designed, trained, and tested to work with text chunks of about 1000 symbols. So, their accuracy is not guaranteed for text chunks that may differ significantly, and the best choice seems to be removing only short sentences. This idea is limited and not applicable when the text chunk is just a single sentence.

The entire algorithm to analyze the influence of the sentence includes the steps below.

1. The classification of the original chunk of the text (paragraph) without any perturbations, saving its class (etalon class) and output (etalon) prediction.

2. The classification of the text chunk without the current sentence (perturbed chunk) with the particular model.

3. Measuring the difference between perturbed and etalon predictions.

4. If the perturbed classification result and the etalon result are different:

4.1 It means that the explanation is found, and ignoring just the current sentence changes the prediction.

4.2 Exploring all combinations (preserving order) of the sentences from the original chunk, which include the current sentence, but may not include any other sentences, searching for the combination that preserves the etalon class. This confirms that the current sentence can guarantee the etalon class almost alone. We call such a chunk, paired with the perturbed one, a complement text pair (CTP). The number of combinations may be significant when the text paragraph contains even a few sentences, so this step may be stopped immediately after the first CTP is found, or limited to only a few combinations to find the one with minimal length, etc.

The second sampling strategy was the removal of n-grams for n = 1..5 (formed without grams intersection in the scope of the single sentence) in the same testing pipeline as sentences.

6. Results and discussion

The numerical evaluation included calculating the explainability index and searching for a complementary text pair (CTP) during the classification of the particular text paragraph. As the main problem of all AIG content classifiers and detectors is false positives: they often detect human-written text as AIG, especially for non-native speakers, we were interested in the explainability of the text pieces that were classified as AI-generated by the average decision of all 5 nets, with the probability greater than 0.75.

According to [17], all ANNs were trained on the dataset in a relatively narrow IT domain in Ukrainian, so we evaluated the explainability using one of our unpublished drafts of the research from 2020 about the application of shallow convolutional neural networks for decision-making.

We modified the initial document by replacing paragraphs in various sections with AI-generated ones. We used the chatbot GUI of ChatGPT (GPT5), Gemini 2.5 Flash, and Claude Sonnet 4.5, applied the same prompt “напиши подібний до наступного фрагменту текст” (English: “write a text similar to the following fragment”), and replaced some paragraphs.

Thus, we obtained an initial human-written document (18 pages in total) and three versions of it with AI-generated pieces in comparable form.

The entire processing of the document included:

– splitting text into pieces of the appropriate size (about 1000 symbols) as all ANNs were trained for this size;

– classification of each piece with all 5 networks, averaging their outputs;

– if the average decision is greater than 0.5, it means the text is classified as AI-generated, and we are interested in cases when the decision is greater than 0.75; we didn’t control whether the classification result is correct according to ground-truth, but we remember that the accuracies of all these networks are about 0.88;

– each such potentially AI-generated chunk was processed with the explanation module, which perturbed the text to explore how the particular ANN model reacts to changes, searching for the influence of every sentence, calculated explainability index, and CTP (if available);

– all results were combined into an HTML web page for further visual investigation.

An example of an explanation representation is shown in Fig. 2.

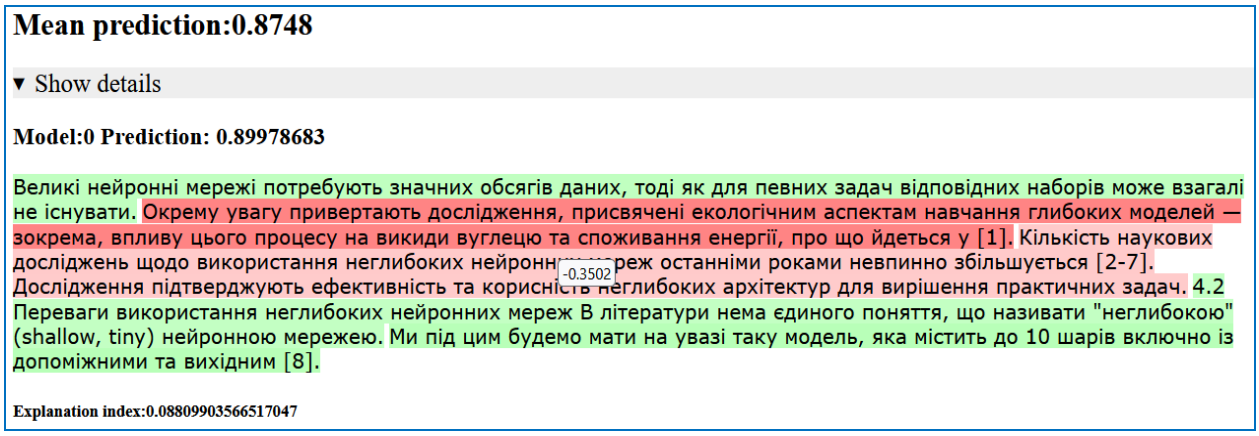

Fig. 2. Example of explanation representation

One can see the mean prediction calculated from the outputs of all 5 nets participating in the processing. The first model (having ID=0) classified the text sample as AI-generated with a probability of about 0.9. The text sample follows next. Different sentences are highlighted in gradient red and green, depending on whether removing the sentence increases (green) or decreases (red) the neural network output. If the prediction decreases after removing it, it means that the decision becomes closer to 0, which corresponds to a human-written class, so the sentence

being removed votes for the AI-generated prediction. Hovering the mouse over each sentence shows the change in the overall classification decision. In this example, removing the second sentence reduced the network's prediction by -0.3502, which is why it has a strong red background. Finally, the explanation index that is about 0.088 for this example is shown under the text fragment. The decomposition of this index according to (1) shows that there are 6 sentences in the fragment, their lengths are 127, 190, 120, 106, 147, and 109 symbols respectively, and the overall length with trailing spaces is 804. The absence of each sentence in the paragraph changes the classification result by such absolute values: 0.058, 0.3502, 0.0056, 0.0039, 0.0425, 0.095, so just the second sentence has a significant influence on the model's output. Summing up all these values with respect to their inverse length coefficients makes the overall normalized explanation index approximately 0.088.

This confirms that the explanation index proposed by (1) is strict enough and may be revised, e.g., taking into account only the sentence with the greatest impact. Its high value shows that there are some short parts of the signal (sentences, words) that affect the ANN prediction.

The example of the CTP is shown in Fig. 3. The sentence on the red background is marked as bold; it means its removal is so significant that it flips the overall classification results, reducing it by 0.4316 from 0.9095. Now the complementary pair for this text is shown below: it still contains this important second sentence and preserves the initial classification prediction (output is greater than 0.99), but without the last sentence of the text chunk. There are multiple such complementary pairs without some other sentences, but all of them contain the sentence in bold and confirm its importance for the decision.

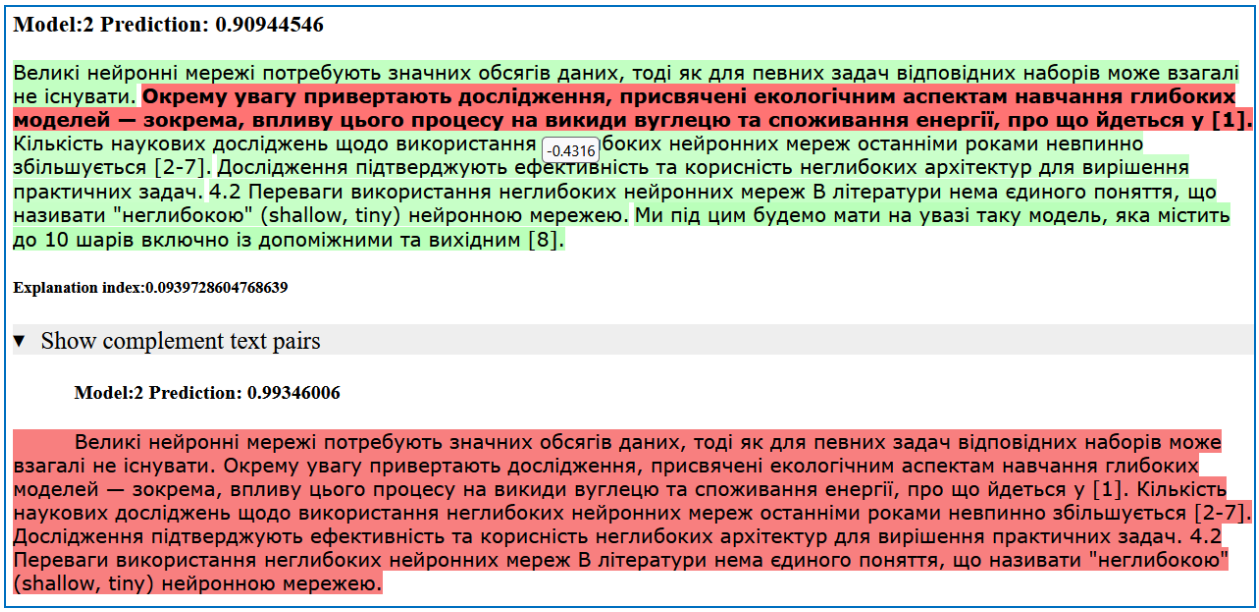

Fig. 3. Example of explanation with CTP

Let's evaluate the qualitative properties for sentence-based and n-grams sampling.

During all experiments, the 9 chunks were classified as AI-generated (the average output by 5 models was greater than 0.75) for the document with GPT 5 insertions, 7 chunks for the document with Gemini 2.5 Flash, and only 6 pieces for the document modified with Claude Sonnet 4.5. For each model, we searched for the explanations, making thus 5x9 = 45 attempts for GPT, 35 for Gemini, and 30 for Claude, respectively. Taking into account that the dataset all models were trained on was created using GPT 4o-mini, the explanations also seem to be more sensitive to GPT-generated fragments.

We were able to find CTP during the hiding of sentences only in 4 cases out of 45 (2 chunks of the text for 2 models) for a document with GPT pieces, 2 different paragraphs for the same model for Gemini (out of 35), and just 1 case for the test with Claude.

The experiments with searching for CTP when hiding n-grams for n = =1..5 were successful only in 1 same case for Model 5 (ID=4) and documents with Gemini-generated pieces. As one can see from Fig. 4, here is the case when just removing of one word (e.g., first one – "Цей") dramatically reduces the classification result by value 0.313 and changes the output class for this chunk from AI-generated class to human-written. These results confirm that finding CTP is more a lucky case for short sequences like words and sentences.

Let's look at the explainability measured according to (1) for different models and text chunks generated by GPT 5, Gemini 2.5 Flash, and Claude Sonnet 4.5 in our test document.

The Table 1 contains the average accuracy (AA) value and average explainability (AE) for the text chunks recognized as AI-generated (output >0.75) per model for each LLM. The correlation coefficient between average output and average explainability is -1 for text chunks generated with GPT, and Gemini, and -0.8 for Claude LLM, meaning the more confident model is on prediction the less it is on explainability (in average) measured according to (1).

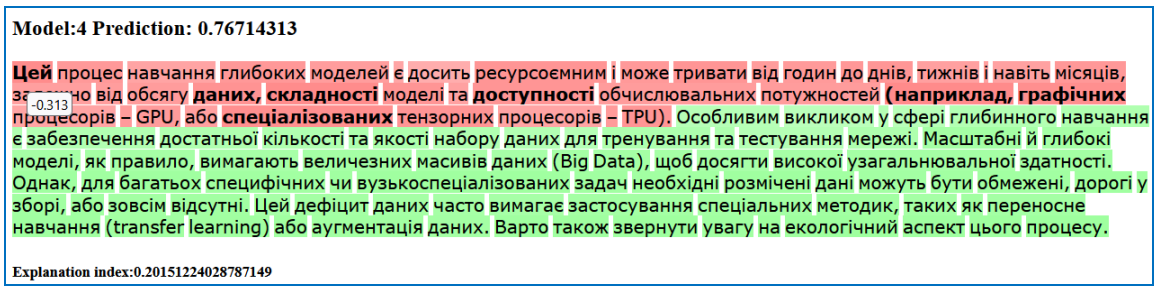

Fig. 4. Example with removing two words

Table 1 – **Average accuracy and explainability values for different models and LLMs**

	Text with GPT 5 AI		Text with Gemini 2.5 Flash		Text with Claude Sonnet 4.5	
Model	AA	AE	AA	AE	AA	AE
1	0.872	0.064	0.863	0.069	0.867	0.064
2	0.886	0.058	0.854	0.075	0.897	0.052
3	0.867	0.066	0.842	0.079	0.891	0.054
4	0.879	0.062	0.893	0.058	0.863	0.076
5	0.803	0.085	0.828	0.082	0.841	0.068

The same negative correlation (<=-0.8) also holds across all LLMs and models when explanations were generated by removing single words, 2-grams, 3-grams, 4-grams, and 5-grams. But taking into account that the explainability index includes the size of the perturbed chunk, and n-grams are shorter compared to sentences, the indices are higher for n-gram sampling. They vary between 0.075 and 0.13, and model 5 was the most explainable for the test document containing GPT insertions, model 3 leads the explainability rating for text with Gemini fragments, and finally, models 1 and 4 have close leading explainability indices for the test documents containing Claude-generated pieces. So, the overall explainability landscape is complicated and may vary from model to model and from sample to sample.

Conclusions

The results presented in the research are ambiguous, highlighting the complexity of AI-generated content classification and the decision-making problems it raises.

We considered searching for the explanations of the classification results for our own shallow ANN models, which are used as black boxes to classify the text chunks in Ukrainian in the IT domain. The explanation module utilizes the perturbation-based idea and modifies the initial text fragment by removing the separate sentences, separate words, and non-intersecting n-grams, $n = \overline{1,5}$.

The experimental modeling showed that finding complementary text pairs is a complex and rare case, and even rarer for n-grams. It is possible to find the contribution of the particular sentence (n-grams) into the overall classification result, but there were just a few cases in our experiments when this contribution was significant enough to change the output class. Probably, the exhaustive search of all combinations of sentences/n-grams could show different results, but this was not the topic of this paper.

We have proposed an explainability index that values the removal of short sentences (n-grams) over long ones and accounts for the significance of this perturbation on the overall classification result.

It would be interesting to investigate not removing sampling but replacing words and/or sentences with synonyms. However, smooth replacement requires a powerful Ukrainian language model, so it could be a possible topic for further work.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Acknowledgements

We dedicate this paper to the Armed Forces of Ukraine and everyone who stands with Ukraine.

The work was supported by the Ministry of Education and Science of Ukraine in the scope of project 0126U002246.

REFERENCES

1. P. Fantozzi and M. Naldi, “The Explainability of Transformers: Current Status and Directions,” *Computers*, vol. 13, no. 4, p. 92, 2024. doi: https://doi.org/10.3390/computers13040092
2. A. Ali, T. Schnake, O. Eberle, G. Montavon, K. R. Müller, and L. Wolf, “XAI for Transformers: Better Explanations through Conservative Propagation,” *Proc. Machine Learning Research (PMLR)*, vol. 162, 2022, pp. 436–451. [Online]. Available: https://proceedings.mlr.press/v162/ali22a/ali22a.pdf
3. A. Dugăeșescu and A. M. Florea, “Evaluation and analysis of visual methods for CNN explainability: a novel approach and experimental study,” *Neural Computing and Applications*, vol. 37, no. 20, p. 14935-14970, 2025. doi: https://doi.org/10.1007/s00521-025-11282-7
4. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2921-2929, 2015. doi: .1109/ https://doi.org/10CVPR.2016.319

5. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618-626, doi: https://doi.org/10.1109/ICCV.2017.74
6. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” *2017 International Conference on Machine Learning*, vol. 70, p. 3319 – 3328. doi: https://doi.org/10.5555/3305890.3306024
7. A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” *2017 International Conference on Machine Learning*, vol. 70, p. 3145–3153. doi: https://doi.org/10.48550/arXiv.1704.02685
8. M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, p. 1135–1144. doi: https://doi.org/10.1145/2939672.2939778
9. S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model Predictions”, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774, doi: https://doi.org/10.48550/arXiv.1705.07874
10. D. Mardaoui and D. Garreau, “An Analysis of LIME for Text Data,” *International Conference on Artificial Intelligence and Statistics AISTATS 2021*, vol. 130, doi: https://doi.org/10.48550/arXiv.2010.12487
11. A. Aghababaei, J. Nikadon, M. Formanowicz, M. Bettinsoli, C. Cervone, C. Suitner and T. Erseghe, “Application of integrated gradients explainability to sociopsychological semantic markers,” Available at: https://arxiv.org/pdf/2503.04989
12. E. Mendez Guzman, V. Schlegel, and R. Batista-Navarro, “From outputs to insights: A survey of rationalization approaches for explainable text classification,” *Frontiers in Artificial Intelligence*, vol. 7, 2024. doi: https://doi.org/10.3389/frai.2024.1363531
13. M. Saarela and V. Podgorelec, “Recent Applications of Explainable AI (XAI): A Systematic Literature Review,” *Applied Sciences*, vol. 14, no. 19, p. 8884, 2024. doi: https://doi.org/10.3390/app14198884
14. B. Wei and Z. Zhu, “ProtoLens: Advancing Prototype Learning for Fine-Grained Interpretability in Text Classification,” *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 4503–4523. doi: https://doi.org/10.18653/v1/2025.acl-long.226
15. H. Yan, L. Gui, and Y. He, “Hierarchical Interpretation of Neural Text Classification,” *Computational Linguistics*, vol. 48, no. 4, p. 987–1020, 2022. doi: https://doi.org/10.1162/coli_a_00459
16. H. Moraliyage, G. Kulawardana, D. De Silva, Z. Issadeen, M. Manic and S. Katsura, “Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers,” *Applied System Innovation*, vol. 8, no. 1, p. 17, 2025. doi: https://doi.org/10.3390/asi8010017
17. O. Peredrii, “Shallow ANN models to classify Ukrainian AI-generated text,” *Control, Navigation and Communication Systems*, no. 4(82), 2025, pp. 108–113. doi: https://doi.org/10.26906/SUNZ.2025.4.108-113
18. O. Gorokhovatskyi, O. Peredrii, and O. Teslenko, “Multiple recursive division explanations for image classification problems,” *Advanced Information Systems*, vol. 9, no. 3, 2025, pp. 5–13. doi: https://doi.org/10.20998/2522-9052.2025.3.01
19. O. Gorokhovatskyi and O. Peredrii, “Recursive Division Explainability as a Factor of CNN Quality,” *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*, vol. 219, 2024, pp. 308–325. doi: https://doi.org/10.1007/978-3-031-70959-3_16

Received (Надійшла) 25.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Передрій Олена Олегівна – кандидат технічних наук, доцент кафедри інформатики та комп’ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Olena Peredrii – PhD, Associate Professor, Department of Informatics and Computer Technology, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: olena.peredrii@hneu.net; ORCID Author ID: https://orcid.org/0000-0003-0390-1931;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57202751577.

Гороховатський Олексій Володимирович – кандидат технічних наук, доцент кафедри інформатики та комп’ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Oleksii Gorokhovatskyi – PhD, Associate Professor, Department of Informatics and Computer Technology, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: oleksii.gorokhovatskyi@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-3477-2132;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=23099879900

Пояснюваність мілких моделей класифікації тексту, згенерованого ШІ, через видалення частин

О. О. Передрій, О. В. Гороховатський

Анотація. У цій роботі ми розглядаємо проблему пояснюваності результату класифікації штучними нейронними мережами (ШНМ) текстових фрагментів, згенерованих ШІ та написаних людиною, у текстах українською мовою в IT-домені. Метою є дослідити, чи можуть модифікації текстових фрагментів на основі пертурбацій, які включають видалення речень, слів та словосполучень, бути корисними для пошуку пояснень. Ми використали п'ять мілких моделей ШНМ (із середньою точністю близько 0.88) і протестували їх на вибірці документів, що містять як написані людиною тексти, так і згенеровані ШІ-фрагменти, створені за допомогою GPT-5, Gemini 2.5 Flash і Claude Sonnet 4.5. Експериментальне моделювання показало, що не завжди можна знайти окреме речення або слово, яке змінює результат класифікації. Ми запропонували індекс пояснювальності, який вимірює загальний вплив усіх змінених зразків на результат класифікації, враховуючи, що короткі пертурбації є більш цінними.

Ключові слова: пояснювальність, «чорний ящик», мілка ШНМ, пертурбація, згенерований ШІ контент, написаний людиною контент, текстовий фрагмент, класифікація тексту, індекс пояснюваності.