# Team iimasnlp at PAN: Leveraging Graph Neural Networks and Large Language Models for Generative AI Authorship Verification

Notebook for the PAN Lab at CLEF 2024

Andric Valdez-Valenzuela[1,†], Helena Gómez-Adorno[1,†]

[1] *Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México, CDMX, México*

[2] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, Ciudad de México 04510, México*

## Abstract

Large language models (LLMs) have led to a surge in machine-generated content across various platforms, posing a challenge for humans to distinguish between machine-generated and human-written text. To address this, there is an urgent need for automated systems that can identify machine-generated content and mitigate related risks. In response, the PAN@CLEF, in collaboration with the Voight-Kampff Task (ELOQUENT Lab) [1][2][3], proposed the Generative AI Authorship Verification task. This study presents a novel model architecture integrating Graph Neural Networks (GNNs), pre-trained Language Models (LLMs), and stylometric features to classify text documents as human-generated or machine-generated. Our approach employs a two-path structure: the first path processes text documents through data augmentation, transforming them into co-occurrence graphs for GNN processing, while the second path extracts and fine-tunes embeddings using a BERT-BASE model along with stylometric features. These embeddings are combined to enhance classification accuracy and robustness. We also detail our data stratification strategy, which involves augmenting human-generated texts to balance the dataset. The effectiveness of our model is demonstrated through extensive evaluation metrics, achieving superior performance over baseline methods.

## Keywords

Generative AI, Text Graph Representation, Graph Neural Networks, Text2graphAPI, Large Language Models

## 1. Introduction

Large language models (LLMs) have become widely available and easily accessible, increasing machine-generated content across various platforms, including Q&A forums, social media, educational resources, and academic settings. Recent advancements in LLM technology, like ChatGPT and GPT-4, enable these models to produce coherent responses to most user inquiries, making them increasingly attractive for replacing human labor in various applications. However, this accessibility has raised concerns about potential misuse, such as generating fake news, impacting the financial services industry, affecting the legal domain, and causing disruptions in educational settings. Given humans' difficulty distinguishing between machine-generated and human-written text, there is an urgent need to develop automated systems capable of identifying machine-generated content to mitigate the associated risks.

Motivated by these challenges, the PAN@CLEF [1][2] proposed a Generative AI Authorship Verification task [3] in collaboration with the Voight-Kampff Task (ELOQUENT Lab). With years of experience in a related but much broader field (authorship verification), they set out to answer whether this task can be solved, starting with the simplest arrangement of a suitable task setup: Given two texts, one authored by a human, one by a machine: pick out the human.

*CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France*

[†] Authors contributed equally

andric_valdez@comunidad.unam.mx (A. Valdez-Valenzuela)

0000-0000-0000-0000 (A. Valdez-Valenzuela); 0000-0000-0000-0000 (H. Gómez-Adorno)

**Table 1**
Total number of problems for Train and validation sets, for human and machine classes.

| Split | *Text Documents* | *Human Texts* | *Machine Texts* |
|---|---|---|---|
| **Training Set** | 23,433 | 12,129 | 11,304 |
| **Validation Set** | 5,654 | 2,827 | 2,827 |

## 2. Background

In recent years, shared tasks focused on automatically detecting AI-generated text have risen, including the prominent Autextification challenge [4], which targets identifying text generated by models in English and Spanish. Notable research from the Autextification-2023 competition includes: "I've Seen Things You Machines Wouldn't Believe: Measuring Content Predictability to Identify Automatically Generated Text" [5]. This system excelled in subtask 1 (differentiating human- and machine-generated text) by evaluating text "predictability" through grammatical accuracy, word frequency, and linguistic patterns alongside a fine-tuned language model representation. Another remarkable work, "Generative AI Text Classification using Ensemble LLM Approaches" [6], achieved top performance in subtask two using an ensemble neural model. It combined probabilities from various pre-trained language models as features for a traditional machine learning classifier, ranking first in both English and Spanish categories.

Another exciting challenge task, SemEval-2024 Task-8 [7], proposed three subtasks over two paradigms of text generation: (1) full text when a considered text is entirely written by a human or generated by a machine and (2) mixed text when a machine-generated text is refined by a human or a human-written text paraphrased by a machine. These three subtasks are composed in the following way: Subtask A is a binary classification task that focuses on identity if a given text was written by a human or a machine; it is split into monolingual (English) and multilingual (Arabic, Russian, Chinese, etc). Subtask B is a multi-class classification task identifying which specific LLM generates a given text among six known options: Human-made, ChatGPT, Cohere, DaVinci, Bloomz, and Dolly. Finally, Subtask C, given a mixed text, where the first part is human-written and the second part is machine-generated, determines the boundary where the change occurs.

## 3. System Overview

This section describes the system overview of our approach: Model Architecture, Data Stratification, and Graph Representation. Model Architecture lays out the structure and detailed workings that drive our method. Data Stratification shows the partition and data augmentation process for the data. Graph Representation explains the text-to-graph representation process.

### 3.1. Data Stratification

The dataset consists of a Training Set and a Validation Set for model training and evaluation (see Table 1). The Training Set includes 23,433 text documents, with 12,129 human and 11,304 machine-generated texts. The Validation Set contains 5,654 text documents, evenly split between 2,827 human and 2,827 machine texts. Initially, there was a significant imbalance, with 14 machine-generated texts for every human-generated text. To balance the dataset, text augmentation was applied to the human texts using the back-translation method with a Large Language Model [1], along with techniques such as word insertions, synonyms, substitutions, and deletions [2]. This augmentation ensured a balanced dataset, facilitating more effective training of models to distinguish between human and machine-generated content.

[1] HuggingFace Models used: Helsinki-NLP/opus-mt-en-ROMANCE to translate from English to another language (Spanish, French, etc.) and Helsinki-NLP/opus-mt-ROMANCE-en to back translate to enlish

[2] We used a python library called *TextAttack*: https://pypi.org/project/textattack/0.0.3.1/

### 3.2. Model Architecture

Figure 1 illustrates a dual-path architecture that classifies text documents as human-generated or machine-generated. This architecture effectively integrates Graph Neural Networks (GNN) with pre-trained Language Models (LLMs) and stylometric features to enhance classification accuracy and robustness.

The top path of the architecture employs a Graph Neural Network (GNN) approach [8]. It begins with the input of text documents, which first undergo data augmentation specifically aimed at human-generated texts (see section 3.1). This augmentation process helps balance the dataset and improve the model's ability to differentiate between human and machine-generated content. The augmented texts are then transformed into co-occurrence graphs using the text2graphAPI [9][3]. In these graphs, nodes represent words, and edges represent words' co-occurrence, capturing the text's structural relationships. Following this transformation, node features are initialized, utilizing the contextualized word embeddings extracted from the fine-tuned BERT LLM to provide a rich text representation [10]. These co-occurrence graphs are subsequently processed by a GNN with a TransformerConv layer [11], which generates graph document embeddings. These embeddings capture the complex relationships and patterns within the text, which are then fed into a classification-dense network in combination with stylometric features. This final dense network is responsible for classifying text documents as human-generated or machine-generated based on the learned features.

The bottom path of the architecture leverages LLMs and stylometric features to complement the GNN approach. Like the top path, text documents undergo data augmentation, with specific techniques applied to human texts to ensure a balanced dataset. In addition, stylometric features are extracted, composed of linguistic and stylistic elements unique to each text. The stylo feature extracted includes mainly text-document stats such as: mean word length, mean sentence length, and the standard deviation of sentence length; furthermore, includes the counter of different punctuation marks such as commas, semicolons, quotes, exclamations, dashes, etc; finally, includes some connector word such as: and, buts, however, mores, this, etc. As an output, it generates a normalized stylometric feature vector that contains all these metrics for each text document from the corpus.

Concurrently, a pre-trained BERT-BASE model is fine-tuned on the text data, and CLS tokens are extracted to obtain document-level embeddings. These embeddings provide a comprehensive representation of the text at the document level. The architecture combines these CLS document embeddings with the previously extracted stylometric features. These combined features are concatenated with the GNN embeddings to form a unified representation that captures both the high-level semantic information and the detailed stylistic nuances of the text. This concatenated embedding is fed into another dense network, which performs the final classification into human-generated or machine-generated categories.

Finally, for each test case (pair of text: text1 and text2), it is required to output the ID of the input text pair and a confidence score between 0.0 and 1.0. A score < 0.5 means that text1 is believed to be human-authored. A score > 0.5 means that text2 is believed to be human-authored. A score of exactly 0.5 means the case is undecidable. To solve this within our architecture, as an output of our classification layer, a soft score probability is obtained for the documents to belong to each class (between 0.0 and 1.0); if text1 is human-authored, apply the rule: 1 - prob_text1; if text2 is believed to be human-authored, apply the rule: prob_text2; lastly, if the prediction is exactly 0.5 (or around an epsilon of 0.01) leave the score as 0.5 which means the problem is indecidable.

### 3.3. Graph Representation

For the graph representation, We used the Co-Occurrence graph, where the words are represented as a node, and the Co-Occurrence of two words within the text document is defined as an edge between the words/nodes. As attributes/weights, edges have the frequency of co-occurrences between words in

[3] https://pypi.org/project/text2graphapi/

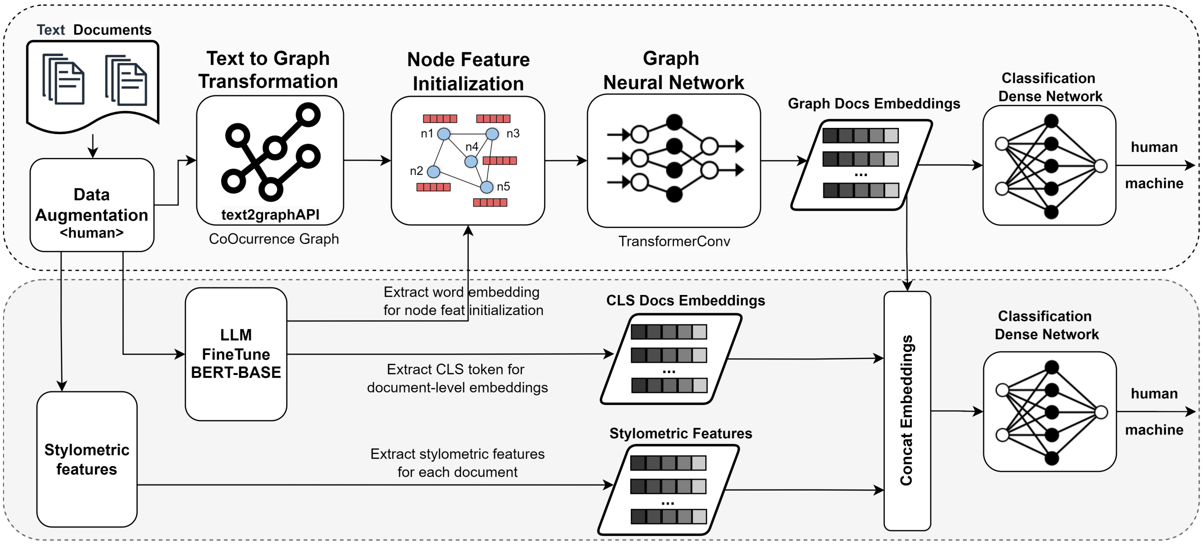
Text Documents
Data Augmentation <human>
Text to Graph Transformation
text2graphAPI
CoOcurrence Graph
Node Feature Initialization
n1
n2
n3
n4
n5
Graph Neural Network
TransformerConv
Graph Docs Embeddings
Classification Dense Network
human
machine
Extract word embedding for node feat initialization
LLM FineTune BERT-BASE
CLS Docs Embeddings
Extract CLS token for document-level embeddings
Concat Embeddings
Classification Dense Network
human
machine
Stylometric features
Stylometric Features
Extract stylometric features for each document

**Figure 1:** Model Architecture.

the text document, and the point-wise mutual information (PMI) measure between each word-to-word relation [4]. As output, we will have one graph representation for each text document in the corpus.

For instance, let's take the following sentence (from one of the documents in the corpus): *millions in Texas lose power as the winter storm falls to -22c* , represented in Figure 2. Each node in the graph corresponds to a unique word from the text, such as "power," "lose," "texas," etc. These nodes are connected by edges, which signify that the words co-occur within the same context or proximity in the text (window size of 2 in this example). The frequency weight indicates how often the connected words appeared together in the document. For example, an edge labeled "freq: 2" means that the two words it connects appeared together twice. The PMI weight quantifies the strength of association between two words, indicating how often the words co-occur more than would be expected by chance. A higher PMI value denotes a stronger association. For instance, PMI values such as 6.08 or 7.40 suggest a significant contextual relationship between the word pairs, even if they do not appear together frequently. This metric helps highlight meaningful word associations that might not be immediately obvious from frequency alone.

Finally, we use the text graph representations for each document and apply an LLM-based node feature initialization process. Subsequently, a GNN processes these graphs to obtain representations that encapsulate semantic meanings, syntactic structures, and contextual information.

## 4. Results

Table 2 presents the evaluation results submitted to the TIRA system [12] for the test test. Different approaches to the task were assessed using several performance metrics: ROC-AUC, Brier Score, C@1, F1 Score, F0.5, and the Mean score. The approaches include various experimental runs and baseline methods evaluated across these metrics to determine their effectiveness. The first section of the table shows our submitted approaches (derived from the architecture presented in the System Overview section), the middle part shows the baselines proposed to the PAN organization, and the last section reports the minimum, median, maximum, 25th, and 75th percentiles over all submissions to the task. For our submission, the approach mark with *(**)* corresponds to the top path of the architecture proposed (using only GNN and stylo features), and, the approach mark *(*)* corresponds to the bottom path (using GNN, LLMs and stylo features).

[4] PMI measures the strength of association between two words by comparing their joint probability to the product of their probabilities

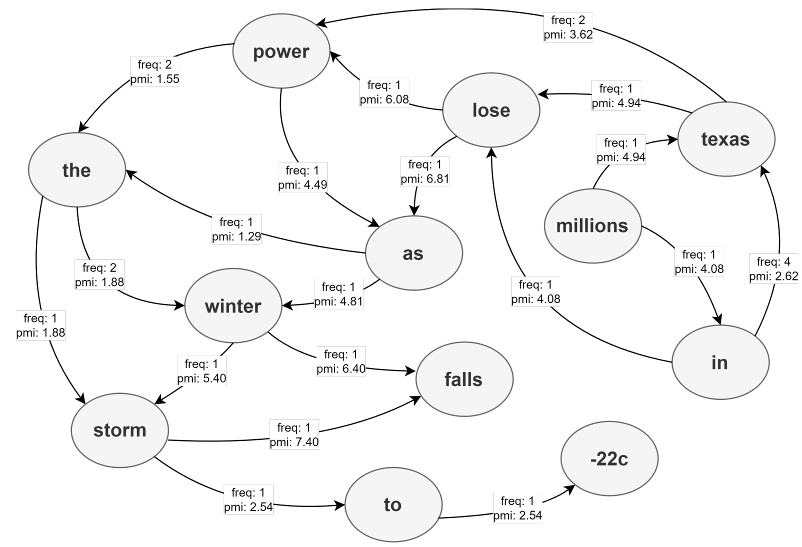

**Figure 2:** CoOccurrence Graph for the text: *millions in texas lose power as the winter storm falls to -22c*

**Table 2**
Overview of the accuracy in detecting if a text is written by an human in task 4 on PAN 2024 (Voight-Kampff Generative AI Authorship Verification). We report ROC-AUC, Brier, C@1, $F_1$, $F_{0.5u}$ and their mean.

| Approach | ROC-AUC | Brier | C@1 | $F_1$ | $F_{0.5u}$ | Mean |
|---|---|---|---|---|---|---|
| final-run4-gnnllm_llmft_stylofeat-partitionB* | 0.992 | 0.992 | 0.992 | 0.992 | 0.991 | 0.992 |
| final-run7-gnnllm_llmft_stylofeat-fullpartitionA* | 0.994 | 0.987 | 0.989 | 0.989 | 0.989 | 0.99 |
| final-run10-gnnllm_llmft_stylofeat-fullpartitionB* | 0.989 | 0.984 | 0.987 | 0.987 | 0.988 | 0.987 |
| final-run8-gnnllm_stylofeat-fullpartitionB** | 0.975 | 0.972 | 0.975 | 0.975 | 0.975 | 0.974 |
| final-run6-gnnllm_llmft_stylofeat-fullpartitionA* | 0.971 | 0.97 | 0.971 | 0.971 | 0.967 | 0.97 |
| final-run9-gnnllm_llmft_stylofeat-fullpartitionB* | 0.97 | 0.97 | 0.97 | 0.971 | 0.959 | 0.968 |
| Baseline Binoculars | 0.972 | 0.957 | 0.966 | 0.964 | 0.965 | 0.965 |
| Baseline Fast-DetectGPT (Mistral) | 0.876 | 0.8 | 0.886 | 0.883 | 0.883 | 0.866 |
| Baseline PPMd | 0.795 | 0.798 | 0.754 | 0.753 | 0.749 | 0.77 |
| Baseline Unmasking | 0.697 | 0.774 | 0.691 | 0.658 | 0.666 | 0.697 |
| Baseline Fast-DetectGPT | 0.668 | 0.776 | 0.695 | 0.69 | 0.691 | 0.704 |
| 95-th quantile | 0.994 | 0.987 | 0.989 | 0.989 | 0.989 | 0.990 |
| 75-th quantile | 0.969 | 0.925 | 0.950 | 0.933 | 0.939 | 0.941 |
| Median | 0.909 | 0.890 | 0.887 | 0.871 | 0.867 | 0.889 |
| 25-th quantile | 0.701 | 0.768 | 0.683 | 0.657 | 0.670 | 0.689 |
| Min | 0.131 | 0.265 | 0.005 | 0.006 | 0.007 | 0.224 |

The six experimental approaches showcase strong performance overall. The **final-run4-gnnllm_llmft_stylofeat-partitionB** stands out with a consistent score of 0.992 across all metrics, indicating highly reliable results. Similarly, **final-run7-gnnllm_llmft_stylofeat-fullpartitionA** exhibits slightly higher performance in ROC-AUC (0.994) but a marginally lower Brier Score (0.987), with an overall mean score of 0.99. **final-run10-gnnllm_llmft_stylofeat-fullpartitionB** also demonstrates robust performance, achieving a ROC-AUC of 0.989 and a Brier Score of 0.984, resulting in an overall mean of 0.987. Meanwhile, **final-run8-gnnllm_stylofeat** and **final-run6-gnnllm_llmft_stylofeat** show slightly lower performances with mean scores of 0.974 and 0.97, respectively.

Regarding the baselines, **Binoculars** perform comparably well, achieving an overall mean of 0.965 with a strong ROC-AUC (0.972) and F1 score (0.966). But, all of our approaches outperformed the baseline score reported.

**Table 3**
Overview of the mean accuracy over 9 variants of the test set. We report the minumum, median, the maximum, the 25-th, and the 75-th quantile, of the mean per the 9 datasets.

| Approach | Minimum | 25-th Quantile | Median | 75-th Quantile | Max |
|---|---|---|---|---|---|
| final-run4-gnnllm_llmft_stylofeat-partitionB | 0.561 | 0.825 | 0.958 | 0.990 | 0.997 |
| final-run7-gnnllm_llmft_stylofeat-fullpartitionA | 0.798 | 0.923 | 0.966 | 0.986 | 0.991 |
| final-run10-gnnllm_llmft_stylofeat-fullpartitionB | 0.732 | 0.917 | 0.963 | 0.988 | 0.999 |
| final-run8-gnnllm_stylofeat-fullpartitionB | 0.781 | 0.939 | 0.966 | 0.975 | 0.986 |
| final-run6-gnnllm_llmft_stylofeat-fullpartitionA | 0.807 | 0.868 | 0.964 | 0.973 | 0.991 |
| final-run9-gnnllm_llmft_stylofeat-fullpartitionB | 0.532 | 0.748 | 0.952 | 0.968 | 0.985 |
| Baseline Binoculars | 0.342 | 0.818 | 0.844 | 0.965 | 0.996 |
| Baseline Fast-DetectGPT (Mistral) | 0.095 | 0.793 | 0.842 | 0.931 | 0.958 |
| Baseline PPMd | 0.270 | 0.546 | 0.750 | 0.770 | 0.863 |
| Baseline Unmasking | 0.250 | 0.662 | 0.696 | 0.697 | 0.762 |
| Baseline Fast-DetectGPT | 0.159 | 0.579 | 0.704 | 0.719 | 0.982 |
| 95-th quantile | 0.863 | 0.971 | 0.978 | 0.990 | 1.000 |
| 75-th quantile | 0.758 | 0.865 | 0.933 | 0.959 | 0.991 |
| Median | 0.605 | 0.645 | 0.875 | 0.889 | 0.936 |
| 25-th quantile | 0.353 | 0.496 | 0.658 | 0.675 | 0.711 |
| Min | 0.015 | 0.038 | 0.231 | 0.244 | 0.252 |

On the other hand, Table 3 shows the summarized results averaged (arithmetic mean) over 10 variants of the test dataset. Each dataset variant applies one potential technique to measure the robustness of authorship verification approaches. In this evaluation, our best run **final-run4-gnnllm_llmft_stylofeat-partitionB** achieved an score of 0.997

Finally, Our submission scores achieved 14th out of 30 on the leaderboard with a ranking score of 0.727 overall test datasets. On the other hand, for part of the test datasets, this score had to be estimated since the system failed to run on short texts (35 words or less).

## 5. Conclusion

This study explores the efficacy of advanced model architectures in differentiating between human-generated and machine-generated text, addressing the growing concern of automated content misidentification in various domains. We implemented an architecture that combines Graph Neural Networks (GNNs) and pre-trained Language Models (LLMs) with stylometric features, showcasing superior performance compared to the baselines proposed.

Our experimental runs, particularly final-run4-gnnllm_llmft_stylofeat-partitionB and final-run7-gnnllm_llmft_stylofeat-fullpartitionA, consistently achieved good scores across multiple evaluation metrics, significantly outperforming all baseline methods. These results affirm the robustness and reliability of our approach in accurately identifying machine-generated text.

Overall, our results highlight the promise of utilizing advanced GNN and LLM architectures and extensive feature sets to tackle the issues arising from the surge in machine-generated content. Future studies could enhance these approaches further and investigate their applicability across various languages and settings. Also, different text graph representations and graph neural network architecture should be tried with the LLM combination.

## Acknowledgments

This paper has been supported by PAPIIT projects IN104424, TA101722, and CONAHCYT CF-2023-G-64. Also, the authors thank CONAHCYT for the computing resources provided through the Deep Learning Platform for Language Technologies of the INAOE Supercomputing Laboratory.

## References

[1] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[2] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[3] J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the Voight-Kampff Generative AI Authorship Verification Task at PAN 2024, in: G. F. N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[4] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

[5] P. Przybyła, N. Duran-Silva, S. Egea-Gómez, I've seen things you machines wouldn't believe: Measuring content predictability to identify automatically-generated text, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023). CEUR Workshop Proceedings, CEUR-WS, Jaén, Spain, 2023.

[6] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, S. Bhattacharya, Generative ai text classification using ensemble llm approaches, arXiv preprint arXiv:2309.07755 (2023).

[7] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, A. F. Aji, N. Habash, I. Gurevych, P. Nakov, Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection, in: Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval 2024, Mexico, Mexico, 2024.

[8] K. Wang, Y. Ding, S. C. Han, Graph neural networks for text classification: a survey, arXiv preprint arXiv:2304.11534 (2023).

[9] A. Valdez, H. Gómez Adorno, Text2graphapi a library to transform text documents into different graph representations, Available at SSRN 4763799 (????).

[10] B. Jin, G. Liu, C. Han, M. Jiang, H. Ji, J. Han, Large language models on graphs: A comprehensive survey, arXiv preprint arXiv:2312.02783 (2023).

[11] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, Y. Sun, Masked label prediction: Unified message passing model for semi-supervised classification, 2021. arXiv:2009.03509.

[12] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.